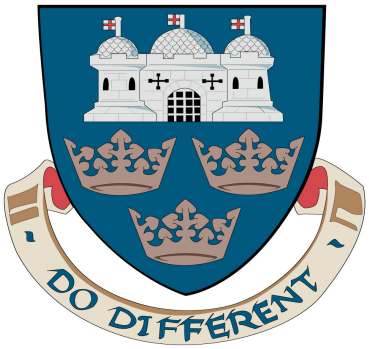
# Reconstruction of intelligible audio speech from visual speech information

Thomas Lee Le Cornu

A thesis submitted for the degree of
Doctor of Philosophy




University of East Anglia
School of Computing Sciences

November 2016

# Abstract

The aim of the work conducted in this thesis is to reconstruct audio speech signals using information which can be extracted solely from a visual stream of a speaker's face, with application for surveillance scenarios and silent speech interfaces. Visual speech is limited to that which can be seen of the mouth, lips, teeth, and tongue, where the visual articulators convey considerably less information than in the audio domain, leading to the task being difficult. Accordingly, the emphasis is on the reconstruction of intelligible speech, with less regard given to quality.

A speech production model is used to reconstruct audio speech, where methods are presented in this work for generating or estimating the necessary parameters for the model. Three approaches are explored for producing spectral-envelope estimates from visual features as this parameter provides the greatest contribution to speech intelligibility. The first approach uses regression to perform the visual-to-audio mapping, and then two further approaches are explored using vector quantisation techniques and classification models, with long-range temporal information incorporated at the feature and model-level. Excitation information, namely fundamental frequency and aperiodicity, is generated using artificial methods and joint-feature clustering approaches.

Evaluations are first performed using mean squared error analyses and objective measures of speech intelligibility to refine the various system configurations, and then subjective listening tests are conducted to determine word-level accu-

racy, giving real intelligibility scores, of reconstructed speech. The best performing visual-to-audio domain mapping approach, using a clustering-and-classification framework with feature-level temporal encoding, is able to achieve audio-only intelligibility scores of 77 %, and audiovisual intelligibility scores of 84 %, on the GRID dataset. Furthermore, the methods are applied to a larger and more continuous dataset, with less favourable results, but with the belief that extensions to the work presented will yield a further increase in intelligibility.

# Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Ben Milner, for his continuous support throughout my studies, and for his patience and encouragement. Furthermore, I extend my thanks to my secondary supervisors, Professor Stephen Cox and Dr Tony Bagnall, for their additional help.

My sincere thanks goes to my family for supporting me throughout my life and in all my endeavours, and to my partner for always being such a happy and wonderful companion. Additionally, I am grateful to all of my dear friends, some of whom I have gained throughout my studies, and others of whom I have known for a great number of years.

# List of PhD publications

Kahn, F., Milner, B., and Le Cornu, T. (under review). Using Visual Speech Information in Masking Methods for Audio Speaker Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Le Cornu, T. and Milner, B. (2015a). Reconstructing Intelligible Audio Speech from Visual Speech Features. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3355–3359.

Le Cornu, T. and Milner, B. (2015b). Voicing classification of visual speech using convolutional neural networks. In *1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP).*

Le Cornu, T. and Milner, B. (in press). Generating Intelligible Audio Speech from Visual Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Websdale, D., Le Cornu, T., and Milner, B. (2015). Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 638–642.

# Contents

# List of Abbreviations

**2D-DCT** Two-dimensional discrete cosine transform

**AAM** Active appearance model

**AMS** Analysis-modification-synthesis

**ASM** Active shape model

**ASR** Automatic speech recognition

**AVSP** Audiovisual speech processing

**AVSR** Audiovisual speech recognition

**CNN** Convolutional neural network

**DB-LSTM** Deep bidirectional long short-term memory

**DNN** Deep neural network

**DRT** Diagnostic Rhyme Test

$f_0$ Fundamental frequency

**GMM** Gaussian mixture model

**HMM** Hidden markov model

**LPC** Linear predictive coding

**LSTM** Long short-term memory

**LVCSR** Large-vocabulary continuous speech recognition

**MAP** Maximum *a posteriori* probability

**MFCC** Mel-frequency cepstral coefficients

**MMSE** Minimum mean squared error

**MRT** Modified Rhyme Test

**MSE** Mean squared error

**PCA** Principal component analysis

**PDF** Probability distribution function

**PEC** Phoneme equivalence class

**PESQ** Perceptual evaluation of speech quality

**ReLU** Rectified linear unit

**RGB** Red-green-blue

**RNN** Recurrent neural network

**SRT** Speech reception threshold

**SSI** Silent speech interface

**STOI** Short-time objective intelligibility

**TTS** Text-to-speech

**VAD** Voice activity detection

**VC** Voicing classification

**VQ** Vector quantisation

**WER** Word error rate

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation and aims

Visual-to-audio speech reconstruction is the process of generating audio speech signals using only visual information that can be extracted from a video of a speaker's face. Two major applications are envisaged for visual-to-audio speech reconstructions: surveillance scenarios where only video footage is available of a target, and silent speech interfaces for laryngectomy patients. To reconstruct an audio speech signal, a model of speech production can be used where the necessary parameters required to drive the model are estimated from visual speech. A typical set of speech production model parameters include source excitation and vocal-tract filter, or spectral-envelope. In this work, it is assumed that the only parameter that can be estimated from visual speech, albeit with some degree of error, is vocal-tract filter information. As only the visual articulators can be seen, it is not possible to obtain fundamental frequency, and it will be difficult to obtain a voicing decision. Thus, the problem is of generating suitable model parameters given only visual speech information. An abstract representation of the problem

the work in this thesis attempts to solve is shown in Figure 1.1.

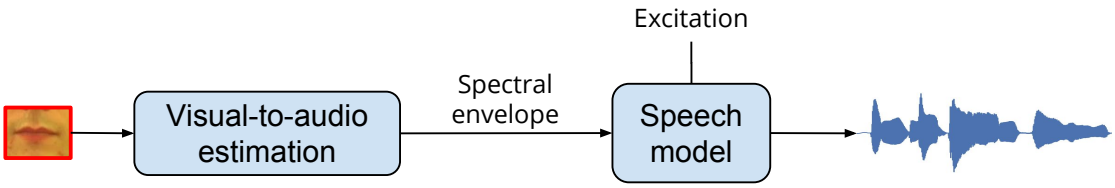



**Figure 1.1:** A general overview of the visual-to-audio speech reconstruction process, where visual information from the mouth of a speaker is used to estimate speech production model parameters to reconstruct audio speech signals.

When a target is being monitored during surveillance, it is often necessary to determine what is being spoken to understand intent and purpose. Accordingly, a microphone would allow for capture of speech signals that can then be processed in a passive manner using an automatic speech recognition system, or by an active listener, to produce transcriptions. However, it may perhaps be the case that using a microphone is infeasible, and the only signal available of the target is a visual stream from a CCTV system or other video recorder. In such a scenario, it is necessary to perform lip-reading, either using automatic systems, or through the expertise of a professional lip-reader. However, automatic lip-reading systems use information about the mouth, lips, and other visual articulators extracted from a visual stream to produce a transcription of the utterance. This process is analogous to automatic speech recognition systems, whereby an audio domain signal is processed to provide a word-level transcription. The work presented in this thesis differs in allowing for audio utterances to be reconstructed directly from visual speech information extracted from the face of a target.

Similarly, visual-to-audio speech reconstruction has application in silent speech interfaces, including for people who have had a laryngectomy. A laryngectomy is an operation where the larynx is removed and, accordingly, the coupling between the lungs and mouth, thus removing the vocal folds and ability to produce speech. Commonly, such people use electrolarynx medical devices, otherwise known as

artificial voice-boxes, to provide a source of voiced excitation. Another solution is to use Permanent Magnet Articulography (PMA) [Hofe et al., 2011; Gonzalez et al., 2015]. In PMA systems, small magnets are placed on the tongue and lips of a speaker, with their locations during articulator movement parameterised using a device containing magnetic sensors. The PMA parameters are subsequently mapped to a set of speech parameters necessary to drive a speech production model. Accordingly, given a sufficiently high audio intelligibility obtained using the work presented in this thesis, it may be possible to construct a device that allows for the processing of visual information with no need to use objects attached to a speaker.

In comparison to audio speech, the primary limitation of visual speech is that it is less effective at conveying information. This occurs as the acoustic speech signal is a consequence of the positions of all the vocal organs in the human speech production system, and the information which can be obtained from the visual stream is limited to that which can be seen of the mouth, lips, teeth, and tip of the tongue [Bernstein, 2012]. The visual modality tends to be used by normal listeners in situations where there are high levels of audio noise, giving clues as to the duration of words and for discerning between audibly confusable phonemes. However, there is evidently sufficient information present in the visual stream to allow for deaf or hard-of-hearing persons to be able to perform lip-reading; although studies show there is a large variation in the abilities of such persons [Lan et al., 2012]. Research investigating correlations between audio and visual speech representations, as used in audiovisual speech processing applications, show that good correlations can be achieved dependent on the feature representations in question [Almajai and Milner, 2007]. It is these correlations between the audio and visual modalities that are exploited in this work.

Due to limitations on the available audio speech information that can be ob-

tained from visual speech, the focus in this work is on producing intelligible audio speech reconstructions with less concern for the quality of the utterances, at least to the extent to which intelligibility is affected. Furthermore, it should be emphasised that the audio speech signal is reconstructed using parameters directly obtained from visual information. This is in contrast to producing audio speech outputs by performing automatic lip-reading on a video signal to derive a word-level transcript, and, therefore, linguistic features, that are subsequently input into a text-to-speech system.

Deriving speech production model parameters from visual information can be considered a domain mapping problem, where it is necessary to produce audio feature estimates from visual features. In addition to visual-to-audio, other domain mappings in speech processing include acoustic-to-articulatory (and vice-versa) for speech coding and speech synthesis applications [Toda et al., 2008], and audio-to-visual for animation requirements [Hong et al., 2002]. Such mappings typically require the parametrisation of the information contained within each domain using particular feature representations, which can then be mapped between using statistical models and machine learning techniques. For this work, it is necessary to consider audio and visual feature representations that exhibit a strong correlation, and, for the audio feature, allow for the reconstruction of speech model parameters. Moreover, having shown application in numerous area of speech processing, Gaussian mixture models and neural networks are explored for use as the domain mapping models.

From Figure 1.1 it can be seen that the two parameters required for the speech production model are spectral-envelope and excitation information. Three approaches are explored in this work for performing visual-to-audio domain mapping to produce spectral-envelope estimates. The first approach investigates various combinations of regression models (Gaussian mixture models and neural networks),

audio features, and visual features, where the audio features are estimated directly from the visual features. The second approach reformulates the problem as one of clustering-and-classification, where vector quantisation techniques are used to produce audio feature codebooks, and the entries are estimated from input visual features using deep neural networks. Furthermore, investigations are conducted on incorporating longer-range temporal information (up to 350 ms in length) at the feature-level, by grouping windows of audio and visual feature vectors. In the third approach, temporal encoding is performed directly at the model-level, where two methods are explored for this task: Viterbi decoding and recurrent neural networks. To produce the excitation information, three artificial methods of fundamental frequency contour generation are proposed, and two methods of obtaining aperiodicity estimates are presented using vector quantisation techniques and convolutional neural networks.

## 1.2 Thesis structure

The remainder of this thesis is organised as follows. The current literature is reviewed in Chapter 2, with the main focus on domain mappings, human speech perception, the intelligibility of speech, and the use of visual speech in myriad areas of speech processing. As the topic of this work is to produce intelligible audio speech reconstructions, it is first necessary to understand what occurs during speech perception and what characteristics of the speech signal make it intelligible. To put this work into context, the use of visual speech in the areas of ASR, speech enhancement, lip reading, and silent speech interfaces is explored.

Chapter 3 provides an overview of the human speech production system and the various components necessary for the process to function. In relation to human speech production, details are given on machine models of speech production, with

a review of commonly used speech reconstruction models, including the model chosen for this thesis, namely STRAIGHT. Furthermore, a number of audio and visual feature representations are described, having shown application in many areas of speech processing.

As a necessary parameter required for speech production models, details about excitation are considered in Chapter 4. Initially, methods for producing artificial fundamental frequency contours are presented, which is then followed by the description of two methods for producing time-frequency aperiodicity surfaces. The first method is based upon using convolutional neural networks as a front-end for performing visual feature extraction, and then to perform classification; and the second method uses a joint clustering of spectral-envelope and aperiodicity estimation using techniques from the area of vector quantisation. The artificial fundamental frequency methods, and the two aperiodicity methods, are evaluated to determine their performance, and the intelligibility results they achieve.

In Chapter 5, the first method for producing spectral-envelope information given input visual features is presented. A number of configurations of different statistical model, audio feature representation, and visual feature representation are explored within a regression framework. Results from subjective listening tests are presented to show the intelligibility of the best configurations for audio-only, audiovisual, and visual-only scenarios. It is found that using deep neural networks results in the greatest overall correlations, despite the intelligibility results being lower than competing configurations.

The use of deep neural networks is explored further in Chapter 6, where they are configured for classification and used within a clustering-and-classification framework for spectral-envelope estimation. Here, the techniques of vector quantisation are used to allow class labels to be assigned to audio feature vectors, which can be

estimated by a deep neural network. Informal listening tests show that there is a marked improvement in intelligibility over the previous regression system. Further improvements to the system are made by incorporating greater temporal information at the feature-level. A number of configurations are evaluated objectively, with the best performing model chosen for further subjective evaluations.

As an extension to the clustering-and-classification approach, work on incorporating temporal information at the model level is presented in Chapter 7. Here, two methods are detailed. Firstly, an approach using the Viterbi algorithm, which although performs well in certain situations, fails to provide sufficiently high objective scores. Secondly, a recurrent neural network using the long short-term memory architecture is used to produce sequences of audio feature estimates from input sequences of visual features. This approach is evaluated objectively, with the best performing configuration explored further using subjective listening tests.

Evaluations of the best performing methods from chapters 5, 6, and 7 are presented in Chapter 8 for two datasets. For the first dataset, subjective intelligibility tests are conducted to evaluate the performance of the three visual-to-audio approaches for obtaining intelligible speech reconstructions. The results from the feature-level and model-level methods show significant improvements over the regression system, and demonstrate that, for the given dataset, high intelligibility can be achieved when reconstructing audio speech from visual information. Evaluations are then presented for the second dataset, which has a larger-vocabulary and less-constrained speech, to determine how the models perform on a bigger dataset. Lower intelligibility is achieved, however, the results suggest that higher intelligibility could be obtained given a more concentrated effort.

Finally, in Chapter 9, the work and results presented in this thesis on intelligible audio speech reconstruction using visual speech are summarised. Additionally, the

limitations of this work are discussed with a number of possible avenues of future work outlined, with a focus on applying the visual-to-audio techniques for two specific applications, and for improving the performance of audiovisual speech enhancement and speaker separation systems.

# Chapter 2

# Literature review

## 2.1 Introduction

This chapter presents a literature review on domain mapping models, audio and visual speech perception, the intelligibility of speech signals including subjective and objective measures, and applications of the visual modality in speech processing. As the aim of this thesis is to perform visual-to-audio mapping, other domain mappings are considered to motivate model selection and design choices. Audio speech perception is considered to determine what makes speech intelligible and how it is affected by modifications to various speech parameters. Visual speech perception is then examined to establish how the visual modality aids normal and hard-of-hearing listeners, especially for speech in the presence of background noise. Objective and subjective measures of speech intelligibility are considered for use in evaluating the performance of different system configurations when producing intelligible audio speech reconstructions. The use of the visual modality in audiovisual and visual-only speech processing applications is explored to motivate design decisions for this work.

Section 2.2 begins with a review of domain mapping models as used in various areas of speech processing. Such mappings consider articulatory and acoustic information for a variety of applications, and audio-to-visual mappings for animation purposes. In Section 2.3, an investigation is conducted on the factors affecting audio speech intelligibility, with an emphasis on audio speech perception and what effect modifications to speech parameters have on intelligibility. In Section 2.4, an overview of subjective and objective intelligibility tests is given. Subjective testing is considered the most appropriate way for determining intelligibility as testing involves real human listeners, however, for convenience, objective measures are often to used in an attempt to predict speech intelligibility. A review of the literature on visual speech perception is presented in Section 2.5, discussing the benefits of the modality for normal and hard-of-hearing listeners in clean and noisy speech conditions. This section is followed by details of speech processing applications using the visual modality for combined audiovisual systems and for visual-only systems in Section 2.6, where visual speech offers numerous benefits including, primarily, robustness to audio noise. Lastly, this chapter is summarised in Section 2.7.

## 2.2 Domain mappings

One of the major components of this work is the estimation of audio information from visual speech, which can be considered as a domain mapping problem. Other examples of domain mappings in speech processing include articulatory-to-acoustic (and vice versa) and audio-to-visual. Accordingly, a review of the domain mapping problem, with reference to estimation models and important considerations, motivates model selection and design choices for this work.

Articulatory-to-acoustic mapping models, and the inverse problem of acoustic-to-articulatory mapping, have application for speech coding, speech synthesis, and

speech modification scenarios [Toda et al., 2008]. The ability to perform the mapping depends on the fact that articulatory information determines the resonant characteristics of the vocal tract during speech production, where the acoustic output is realised due to these configurations. Approaches to this problem include using hidden Markov models to determine articulatory movements from acoustic features with and without the use of phonemic information [Hiroya and Honda, 2004], and using joint probability density Gaussian mixture models with application of the minimum mean square error criterion for performing the conversion [Toda et al., 2008]. The primary difficulty with estimating articulator movements from audio speech is that there exists a one-to-many mapping of speech acoustics to articulatory configurations. Experiments conducted by Atal et al. [1978], using a computer simulation to examine the relationship between vocal-tract configuration and acoustic output, found that different vocal-tract shapes could yield acoustic outputs with nearly identical values for the frequencies and amplitudes of the first three formants. One proposed solution to this one-to-many mapping between the audio and articulatory domains is achieved by incorporating visual speech (modelled using AAM features) yielding audiovisual-to-articulatory models [Katsamanis et al., 2009].

The inverse of the visual-to-audio domain mapping problem is that of estimating the facial movements of visual speech directly from the acoustic waveform. Audio-to-visual mapping models have application in animation, and computer-aided interfaces for tools such as virtual agents, email readers, and so on, where the goal in these applications is to achieve realism. Having shown application for other domain mapping problems, multi-stream hidden Markov models can be utilised to map from acoustic speech features (e.g. MFCCs or LSPs) to visual speech features (e.g. AAMs). In [Fu et al., 2005], an HMM state sequence is determined from input audio using the Viterbi algorithm, and the visual output is

taken as the mean of the Gaussian mixtures corresponding to each state. Other approaches to performing this mapping include using neural networks [Hong et al., 2002; Massaro et al., 1999], and more recently deep neural networks. In a system proposed by Taylor et al. [2016], DNNs are used to estimate a temporal window of facial movement parameters from an input window of acoustic speech features. Averaging of the overlapping estimated visual speech features is performed to generate continuous and smooth speech animations.

Numerous investigations exist detailing the degree of correlation between acoustic and visual speech information (explored further in Chapter 3), however, fewer examples are found of actually using estimated audio information output by a visual-to-audio mapping model. Two areas of speech processing using audio estimates from visual information include speaker separation and speech enhancement. Girin et al. [2001] explore a number of techniques for incorporating visual information into audio speech enhancement systems, where they use linear and non-linear models to produce direct estimates of enhancement filters, and estimates of spectral information from which filter coefficients are derived. In Almajai and Milner [2011], Gaussian mixture models are used to estimate log-filterbank audio features from visual information, which are employed for Wiener filtering of noisy speech signals. Similar ideas are applied in Rivet et al. [2014] and Khan and Milner [2015] for audiovisual speaker separation.

## 2.3 Intelligibility of audio speech

This section begins with a review of the literature pertaining to what makes audio speech intelligible, with emphasis on audio speech perception and the effect of modifications to speech production model parameters. For this work, it will be beneficial to understand what characteristics of audio speech signals contribute

to intelligibility, and what ought to be considered for design decisions and for evaluating the performance of different systems.

The intelligibility of a signal speech is a measure indicating to what extent it is deemed comprehensible. Weismer [2008] states that, for a given situation, speech intelligibility is dependent on the characteristics of both the speaker and the listener, the speech material, and the channel. Accordingly, there are a number of factors that can affect how intelligible a speech signal is. For example, when considering the communication channel, audio signals can be negatively influenced by background noise, and acoustic effects such as reverberation and echo. With regards to speaker characteristics, speech disorders such as dysarthria may result in a speech signal where the intelligibility is impaired at the outset due to poor phonation [Morales and Cox, 2009]. Furthermore, listeners may be afflicted with hearing impairments resulting in poor frequency selectivity, and is a problem further compounded in noisy environments [Baer et al., 1993].

### 2.3.1 Audio speech perception

Audio speech perception, as occurs in humans, functions by merging information extracted from independent frequency regions, to form the sounds units of speech (phonemes) that are then combined to form, syllables, words, and sentences [Allen, 1994]. The frequency regions exist as overlapping frequency bands, also known as critical bands and proposed by Fletcher and Munson [1933], where the independence of the frequency regions is important for enabling humans to continue to recognise speech despite errors in other frequency regions caused by masking. Such errors in an audio signal may be introduced due to the addition of noise and reverberation. The information extracted from each frequency channel is combined to form a feature set to maximise the identification of the correct phoneme [Fletcher,

1953]. For example, if a speech sound is processed by two separate channels, and the upper channel is corrupted by noise, the phoneme information will be entirely extracted from the lower channel.

### 2.3.2 Effect of speech parameter errors on intelligibility

In this thesis, audio features are estimated from visual speech information which are subsequently input into speech production models to produce audible speech reconstructions. As it is unlikely that the speech parameter estimates will be perfect, it is necessary to understand what effect errors in the parameters have on the intelligibility of reconstructed speech signals. The speech parameters considered are spectral-envelope, fundamental frequency, and phase.

It is assumed that only broad spectral-envelope information can be estimated from visual speech to any real degree, where such estimates will exhibit an amount of smoothing. It is, therefore, necessary to understand how intelligibility will be affected by such alterations of the spectral-envelope. When considering fundamental frequency, it will not be possible to determine usable information from the speech signal, and, accordingly, artificial contours will be required as input to speech production models. Therefore, investigations on the intelligibility of speech with $f_0$ modifications will provide valuable information when developing the artificial-$f_0$ methods. Similarly, phase information can not be estimated from visual speech, yet in Paliwal and Alsteris [2003] the contribution to speech intelligibility in humans is found to be equivalent to that of the power spectrum. Studies conducted on speech intelligibility with altered phase information will guide the selection of speech production models in the next chapter.

#### 2.3.2.1 Spectral envelope

Ter Keurs et al. [1992] investigated the effect of spectral smearing on the intelligibility of speech produced by a female speaker in the presence of noise. The smearing of the spectral-envelope, simulating lowering of the frequency resolution in the auditory system, was performed by convolution of the short-term power spectrum with a Gaussian-shaped filter. The effect of smearing on the spectral slope is for it to tend towards a straight line. It was found that there is a direct relationship between the resolution of the spectral-envelope and the intelligibility of speech in the presence of noise for sentence material. Further experiments conducted determined the extent to which vowels and consonants in clean and noisy conditions were affected by spectral smearing. Under noisy conditions the effect of spectral smearing was more significant for vowels than consonants.

Ter Keurs et al. [1993] extended their earlier study to include speech from a male speaker, therefore allowing them to compare their earlier results with a female speaker. They hypothesised that any such differences between the genders could possibly be attributed to the difference in amplitudes and bandwidth of the formant peaks, and the depth of the valleys in between. Their results agreed with the earlier work that showed the effect of spectral smearing on the SRT, but they found that there was no significant difference between the intelligibility of the male and female speakers.

#### 2.3.2.2 Fundamental frequency

The fundamental frequency, $f_0$, of a speech signal is the lowest harmonic produced by the vibrating of the vocal folds. In speech perception, the fundamental frequency provides important linguistic cues, and where dynamic changes in the contour serve to indicate the location of important words in an utterance [Cutler et al.,

1997]. A number of studies have been conducted on the intelligibility of speech utterances with modified $f_0$ contours for both English and other languages [Spitzer et al., 2007; Laures and Weismer, 1999; Wang et al., 2013]. Flattening of the $f_0$ contour, as occurs in some speech disorders resulting in audio utterances that are more monotonic in nature, has a detrimental effect on intelligibility. In a study conducted by Miller et al. [2010], a number of modifications to the contour were performed in addition to flattening. These included exaggeration of the contour, applying inversion, and replacing the contour entirely with a slowly-oscillating sinusoid. Relative to utterances with the original $f_0$ contour, those with exaggerated and flattened contours exhibited relative reductions in key-word recognition accuracies of 13 %, and with a further loss of intelligibility for the inverted and sinusoidal contours showing a relative reduction of 23 %. Furthermore, in a study on global and fine-grained acoustic speaker characteristics, Bradlow et al. [1996] found a similar tendency for increased intelligibility with a wider range of $f_0$, and, additionally, that there was no apparent correlation between increased speech intelligibility and mean fundamental frequency.

#### 2.3.2.3 Phase

Experiments conducted by Shi et al. [2006] explored the effect of phase errors on speech intelligibility. The phase of speech utterances was altered within an analysis-modification-synthesis framework with increasing levels of corruption from perfect phase to entirely random phase. It was found that intelligibility is dependent on both the amount of phase noise, and the signal-to-noise ratio. At low SNRs ($-10$ dB), an absolute increase in word error rate of 39 % results when the phase is completely random as opposed to when there is no phase noise. These results confirm those found by Paliwal and Alsteris [2003], where they discovered that the phase spectrum can significantly contribute to the intelligibility of speech.

## 2.4 Measures of speech intelligibility

The ideal procedure for conducting intelligibility experiments is to perform subjective evaluations. However, it is not necessarily always convenient to perform subjective listening tests as they require preparation of testing materials, organisation of subjects, and time to conduct the experiments. Therefore, although they should not be considered a replacement for subjective evaluations, objective measures of speech intelligibility are commonly used as they are cheaper, easily repeatable, and less time consuming [Schmidt-Nielsen, 1992]. A review of subjective test configurations and of objective measures frequently used in the literature are provided in this section.

### 2.4.1 Subjective measures

When determining the intelligibility of speech signals using subjective testing, scores are typically calculated from the number of correctly identified responses by listeners to phonemes, words (either meaningful or nonsense), or sentences [Steeneken, 2001], and are based on the perception of speech intelligibility as that of understanding the speech material. When deciding upon the type of subjective listening tests to perform, it is important to understand how the tasks evaluated in the tests relate to the use-cases in the real-world, such that results can be extrapolated appropriately. Evaluations such as rhyme tests and using nonsense syllables can be used to determine the amount of acoustic detail pertaining to the phonetic structure of the material, and are highly repeatable despite perhaps being less realistic. Whereas sentence material can be used to provide more realistic scores where listeners are able to rely less on acoustic-phonetic information with greater emphasis placed on context, at the compromise of being less repeatable [Schmidt-Nielsen, 1992].

Test configurations can be either open response where no choices are presented to the listener, or closed response where a selection of choices are offered. Proposed by House et al. [1963], the Modified Rhyme Test (MRT) is a closed response test consisting of fifty separate six-word lists, where each word is of the form consonant-vowel-consonant (CVC) and are similar sounding. A carrier sentence is typically used and the listener is asked to select which word of the six they believe to be correct. Additionally, the MRT allows for the confusions of phonemes to be determined. A similar closed response test is the Diagnostic Rhyme Test (DRT) consisting of 96 rhyming word pairs where no carrier phrases are used. The DRT provides scores for a set of phonemic features in addition to overall intelligibility [Voiers, 1983]. Rhyme tests are suitable for evaluating segmental acoustic-phonetic information where the difference of individual speech sounds are important.

For evaluating suprasegmental cues (e.g. pitch, duration, and intensity) sentence material can be used. In such tests, intelligibility is typically calculated on either a per-word basis (the number of correctly identified words in a sentence), or at the sentence level, where it is necessary for the entire sentence to be correct. Egan [1948] provides a set of phonetically balanced sentences, where intelligibility scores are determined by the number of correct responses of five keywords within each sentence. Other sentence material, again where keyword accuracy is scored, include those presented by Nye and Gaitenby [1973], where the sentences are semantically anomalous but still grammatically correct. The Speech Reception Threshold (SRT) proposed by Plomp and Mimpen [1979] provides intelligibility scores as the minimum signal-to-noise ratio at which a listener can understand fifty per cent of words within a sentence.

### 2.4.2 Objective measures

A number of objective measures have been proposed for evaluating intelligibility of speech material, where assessments are based on the physical parameters of a transmission channel [Steeneken, 2001]. It should be noted that objective measures are used to predict intelligibility, and are unlikely to provide entirely accurate results.

The Articulation Index (AI), proposed by French and Steinberg [1947], assumes that a speech signal is not uniformly distributed in the frequency domain, and that speech intelligibility can be determined based on the sum of contributions from individual frequency channels that are audible to a listener. The Speech Intelligibility Index (SII) [ANSI, 1997] is a refinement to the AI, that is able to account for band-pass limiting and noise. A frequency-weighting function is introduced to assign greater importance to the signal contained within certain frequency ranges. Another traditional measure is the Speech Transmission Index (STI), developed by Steeneken and Houtgast [1980] in response to the authors being required to conduct numerous subjective intelligibility assessments. The method assumes that acoustic speech information is formed of a sequence of temporal modulations, and that any reduction in such modulations, perhaps due to additive noise or reverberation, will result in a reduction of overall intelligibility. Unlike the AI and SII, the Speech Transmission Index is able to account for non-linear distortions including acoustic effects such as reverb and echo.

Two commonly used measures in the literature for assessing speech intelligibility of synthesis systems [Valentini-Botinhao et al., 2011], and speaker separation methods [Tu et al., 2014], are PESQ and STOI. Originally designed for assessing speech quality for narrowband telephony systems, the Perceptual Evaluation of Speech Quality (PESQ) [Rix et al., 2001] method gives results that have been

shown to correlate well with speech intelligibility evaluations for the cases of interfering speakers [Beerends et al., 2004], and background noise [Ma et al., 2009]. The short-time objective intelligibility (STOI) measure [Taal et al., 2010], was designed for evaluating speech material processed by speech enhancement and speaker separation systems.

Fewer measures exist for predicting speech intelligibility where signal distortions do not result from the interference of background noise. Such distortions may occur due to speech parameter modifications, the effects of speech production models, or the processing of signals by hearing-aids [Kates and Arehart, 2014]. The Normalised Frequency-weighted Distortion measure (NFD) proposed by Websdale et al. [2015], is designed to measure the amount of spectral distortion between an original and processed utterance, showing benefits over STOI and PESQ for predicting utterance intelligibility with spectral-smoothing modifications.

A comprehensive review detailing the performance of various traditional and newly-proposed objective measures for predicting subjective speech intelligibility is presented by Ma et al. [2009] for noisy speech scenarios, and by Websdale et al. [2015] for audio reconstructions of spectrally-smoothed speech using artificial fundamental frequency contours. A summary of the results on correlations between subjective intelligibility scores and objective measures as applied to spectrally-smoothed speech are presented in Table 2.1, using work from experiments conducted in Chapter 5 of this thesis. In addition to those reviewed thus far, the investigation also considers measures for determining the quality (Hearing Aid Speech Quality Index, HASQI) and intelligibility (Hearing Aid Speech Perception Index, HASPI) of degraded speech for listeners using hearing aids [Kressner et al., 2013; Kates and Arehart, 2014], other measures for objectively predicting speech intelligibility (normalised covariance matrix, NCM; and coherence speech intelligibility index, CSII) [Holube and Kollmeier, 1996; Kates and Arehart, 2005],

and measures for determining spectral-envelope distortions (log likelihood ratio, LLR; and distance, CEP) [Kitawaki et al., 1988; Quackenbush et al., 1988]. The objective measures that exhibited the greatest correlations with the subjective intelligibility tests results were STOI and NFD.

**Table 2.1:** Correlation, $r$, and standard deviation of the error, $\sigma_e$, between word-level accuracies and objective measure scores, taken from [Websdale et al., 2015].

| Measure | $r$ | $\sigma_e$ |
|---|---|---|
| PESQ | 0.63 | 0.15 |
| LLR | -0.63 | 0.15 |
| CEP | -0.65 | 0.14 |
| NCM | 0.70 | 0.13 |
| AI-ST | 0.44 | 0.17 |
| CSII | 0.22 | 0.18 |
| $\text{CSII}_{\text{high}}$ | 0.24 | 0.18 |
| $\text{CSII}_{\text{mid}}$ | 0.30 | 0.18 |
| $\text{CSII}_{\text{low}}$ | 0.44 | 0.17 |
| STOI | 0.75 | 0.12 |
| $\text{HASQI}_{\text{nonlin}}$ | 0.62 | 0.15 |
| $\text{HASQI}_{\text{lin}}$ | 0.30 | 0.18 |
| $\text{HASQI}_{\text{comb}}$ | 0.58 | 0.15 |
| HASPI | 0.64 | 0.14 |
| NFD | -0.81 | 0.11 |

## 2.5 Visual speech perception

In this section, a review is conducted on the benefits offered by the visual modality in human speech perception. For the work in this thesis, the original video and reconstructed audio signals can be combined to produce audiovisual media, which is expected to achieve higher intelligibility over either the audio or visual

signals on their own. Accordingly, it is important to understand how visual speech information aids listeners.

Before the discovery of the McGurk effect, phonetic perception of speech signals was considered to exist only in the audio domain. The McGurk effect [McGurk and MacDonald, 1976] occurs when visual information conflicting with audio speech information impacts auditory perception. For example, a repetition of the syllable /ga/ in the visual domain, mixed with the syllable /ba/ in the audio domain, produces auditory perception of the syllable /da/ in normal hearing adults. The effect motivated a shift from the prevailing auditory-only models of speech perception to multi-modal models, where information from audio and visual speech is combined. The two modalities are complimentary, in that, whilst audio speech is sufficiently robust for conveying the majority of the information, visual speech aids in identifying the place of articulation, helping to distinguish between audibly confusable phonemes [Sekiyama et al., 2003].

In comparison to the audio speech signal, visual speech is phonetically lacking, although not to the extent that visual-only word recognition is impossible [Bernstein, 2012]. Using only speech gestures of a speaker's lips and face to understand speech is known as lip-reading [Schwartz et al., 2004]. Summerfield [1992] finds that there is a considerable amount of variation in the ability for humans to lip-read, even for those who are considered better than average, with word recognition scores typically reported in the range of 10–70%. This considerable variability can be attributed to numerous causes including the abilities of the lip-readers themselves, the vocabulary of the spoken material, and the person talking. When conducting subjective experiments which require subjects to lip-read, it is necessary to try and control for this large variation in abilities.

Studies have shown that the visual modality offers benefits for degraded au-

ditory signals where speech intelligibility is improved using information from the visible articulators (lips, teeth, and tongue) [Sumby and Pollack, 1954], and non-verbal movements of the head [Munhall et al., 2004]. The greatest contribution of visual speech information is made when the audio speech signal is significantly masked or corrupted, although the benefit of seeing the mouth of a speaker still remains in clean speech. A study conducted by Middelweerd and Plomp [1987] showed that subjects are able to tolerate an extra 4dB of audio noise using the visual modality, where, for sentence material, a rough increase in intelligibility of 10–15% per decibel results. Moreover, Summerfield [1987] describes further benefits of visual speech such as aiding with speaker localisation, and providing additional segmental speech information such as syllable and word boundaries.

## 2.6 Audiovisual speech processing

In the previous section, investigations into the benefits offered by the visual modality were described with reference to human speech perception. In this section, the literature on using visual speech in speech processing applications to improve performance of audio-only systems, and for visual-only systems, is examined.

The first documented use of incorporating visual speech into an automatic speech recognition (ASR) system was presented by Petajan [1984]. Simplistic binary images of the mouth were used to extract geometric features (area, height, width, etc.) to complement audio features in an ASR system investigated for a small-vocabulary isolated-word dataset, where improvements were achieved over an audio-only system. The primary reason for incorporating visual features into ASR systems is that they are robust to audio noise, although benefits are still seen in clean conditions [Glotin et al., 2001]. Experiments conducted by Potamianos et al. [2003] showed that audiovisual speech recognition (AVSR) systems outper-

form audio-only ASR systems over a wide range of conditions with significant reductions in word error rates (WER) at low SNRs. For a large vocabulary continuous speech recognition (LVCSR) task, their best performing system achieved an 8dB gain, with the combined audiovisual system achieving results at 2dB similar to those for the audio-only system at 10dB. Additionally, as has been witnessed in recent years with audio ASR systems, using deep neural network (DNN) models has yielded a further increase in AVSR performance [Noda et al., 2015].

Other examples of exploiting both modalities for speech processing applications include speaker verification [Dean and Sridharan, 2010], speech enhancement, voice activity detection, and speaker separation. In a study conducted by Almajai and Milner [2011], visual information was used to enhance audio speech signals in low SNRs using a Wiener filter system, where incorporating visual information yields significant improvements in perceptual quality of processed utterances [Abdelaziz et al., 2013]. Given the robustness of the visual modality to noise, voice activity detection has benefited from using both modalities, where a weighting mechanism allows the contributions of the audio and visual streams to be controlled depending on the level of interfering audio noise [Almajai and Milner, 2008]. Furthermore, single channel speaker separation system, where a single microphone recording contains a mixture of two or more speakers, have shown improvements over audio-only systems by integrating visual information for both binary mask [Khan and Milner, 2013] and soft-mask [Khan and Milner, 2015] configurations.

Using the visual modality alone gives rise to automatic lip-reading systems, where descriptions of the visual articulators are used to produce phoneme or word-level transcriptions. A study conducted by Lan et al. [2012] evaluated the performance of a machine lip-reading system and that of six professional human lip-readers. When provided with transcriptions and vocabulary of the material, they observed a large variation in the word recognition rates of the human lip-

readers (0–60%). In comparison, the machine lip reading system achieved a word recognition accuracy of 14%, outperforming four of the six expert lip-readers. Recent automatic lip-reading performance on a large-vocabulary (1000 words) corpus achieves word accuracies of 85%, when using DNNs instead of conventional HMMs for a speaker-dependent configuration [Thangthai et al., 2015]. Due to large inter-speaker variability of the visible articulators, speaker-independent AVSP systems have been considerably less successful than speaker-dependent systems. Experiments on speaker adaptation techniques applied to visual features in a lip-reading task by Almajai et al. [2016] yield mean word recognition accuracies of 55% across twelve speakers, a significant improvement on previous work. However, despite these recent improvements, the performance of state-of-the-art speaker-dependent and speaker-independent lip-reading systems is still far lower than that of the best-performing audio ASR systems.

## 2.7 Summary

This chapter presented a literature review of experiments and investigations considered important for the work conducted in this thesis. The domain mapping problem is discussed initially, where a number of models have been explored for performing the mapping including GMMs, HMMs, and DNNs. When considering different domain mapping problems, the objectives for each are distinct. That is to say, for articulatory-to-acoustic mapping the goal is to reduce parameter dimensionality for speech coding applications, and for audio-to visual work the emphasis is on producing realistic mouth and facial animations. However, for this work, the reconstruction of intelligible audio speech is the main aim.

The intelligibility of speech signals is then considered, where intelligibility provides a measure indicating to what extent a speech signal can be understood, and

is affected by a number of factors. Modifications to speech parameters such as smoothing of the spectral-envelope and flattening of the fundamental frequency, important to consider for speech reconstruction, are shown to affect intelligibility, especially under noisy conditions. For measuring speech intelligibility, subjective or objective tests can be performed. In subjective tests, listeners are asked to transcribe responses such as phonemes, words (nonsense or meaningful), or entire sentences, where scores are determined based on the number of correct responses. Whereas objective measures aim to predict intelligibility, with the main benefits being that evaluations can be performed significantly quicker and cheaper than subjective tests. Two commonly used measures within the literature for evaluation of speech synthesis systems include STOI and PESQ, and are used throughout this thesis.

Next, the focus turns to visual speech perception as occurs in humans. The visual modality offers benefits for both normal and hard-of-hearing listeners, providing information regarding the place of articulation to aid with the differentiation of audibly confusable phonemes, the identification of word boundaries, and for speaker localisation. The main contribution of the visual modality occurs under low SNR conditions where there is considerable interfering background noise. Accordingly, such benefits have motivated integrating the visual modality into speech processing applications such as ASR, speech enhancement, and speaker separation. Even under clean conditions, ASR systems have shown improvements when both the audio and visual modalities are combined, with the biggest reduction in WERs occurring in low SNR conditions. Visual-only systems give rise to automatic lip-reading, where word-level transcriptions are produced using only visual speech information. Although WERs are considerably higher than the best performing ASR systems, recent results have shown significant improvements.

In the next Chapter, a review of speech production is presented, initially focus-

ing on how the process occurs in humans, and then on speech production models as used for speech synthesis and reconstruction. This is followed by details of common audio and visual features, and the correlations that exists between them—the relationships of which are exploited in this thesis.

# Chapter 3

# Speech production

## 3.1 Introduction

This chapter begins with an overview of human speech production to determine the important components of the process to motivate design decisions for this work. It is necessary to establish what is required from visual speech information to drive a speech production model, where various models are assessed for producing audio speech reconstructions. Evaluations are conducted on two feature representations of the spectral-envelope, a necessary parameter of speech production models, and two feature representations of the visual articulators, from which spectral-envelope information can be estimated.

The thesis of this work is that an intelligible audio speech signal can be reconstructed using solely visual speech information extracted from a video of a speaker. Given a video focused on the face of a speaker, a sequence of visual feature vectors, $\mathbf{v}_i$, can be extracted, localised about the mouth (lips, teeth, and tongue). This visual sequence is then passed to a visual-to-audio domain mapping model to produce a sequence of corresponding audio feature estimates, $\hat{\mathbf{a}}_i$. The audio sequence

can then be transformed into a time-frequency spectral-envelope, which is used, with the addition of source excitation information, to reconstruct an audio speech signal using a speech production model. Figure 3.1 provides a pictorial overview of this process.

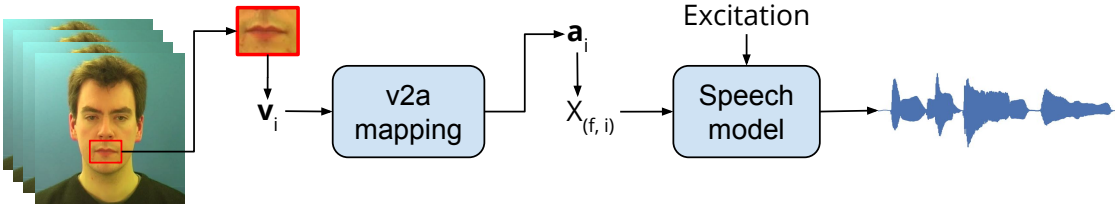



**Figure 3.1:** Audio speech reconstruction system with necessary components. Visual features are extracted from a video of speaker which are then input to the visual-to-audio mapping model outputting audio feature estimates, which are used to produce spectral-envelopes. A speech production model is then used to reconstruct the audio speech output given the spectral-envelope and an artificial excitation signal.

Speech production models have application in speech processing for reconstructing or synthesising audio speech signals given a set of input parameters. A typical set of parameters include spectral-envelope and source excitation information, and are motivated by the mechanisms of the human speech production system. To reconstruct an audio speech signal, the source excitation information, which is either a pulse-train with periodicity pertaining to the fundamental frequency for voiced frames or simply white noise for unvoiced frames, is modulated by a filter that models the resonances of the vocal-tract.

In this work, there is an assumption that sufficient audio speech information can be estimated from visual speech. As discussed in Chapter 1, it is not possible to obtain estimates of the fundamental frequency of a speaker from visual speech as the vocal-cords cannot be seen. Obtaining a voicing decision will also be difficult and depend on what information can be inferred from the shape of the mouth in relation to the speech sound being produced. Therefore, it is assumed that only spectral-envelope information can be estimated adequately from the visual

information. Accordingly, the correlation of audio and visual speech information is discussed, with further details provided for the selected feature representations of each modality.

The remainder of this chapter is organised as follows. In Section 3.2, an overview of the mechanisms for human speech production is given, discussing the main organs and the separation of speech signals into source excitation and vocal-tract filter components. Following on from human speech production, a complementary review of common speech production models is presented in Section 3.3, including details of their workings. Various audio features, for representing spectral-envelope, and visual features, for representing the mouth movements of a speaker, are examined in Section 3.4 and Section 3.5, respectively. Lastly, a summary of this chapter is provided in Section 3.6.

## 3.2 Human speech production

Speech is one of the main forms of human communication, resulting from the processing of ideas and thoughts into words and sentences, which are then vocalised within the human speech production system. Upon reaching the auditory system of a listener (audio speech perception is discussed in Chapter 2), the acoustic speech signal is processed to derive the meaning of what was conveyed. The human speech production system consists of a number of organs (the vocal organs) where the interactions between them result in audio speech signals.

### 3.2.1 Anatomy

The main organs of human speech production are the lungs, larynx, pharynx, nose, the hard and soft palates, tongue, teeth, and lips [Holmes and Holmes, 2001]. These

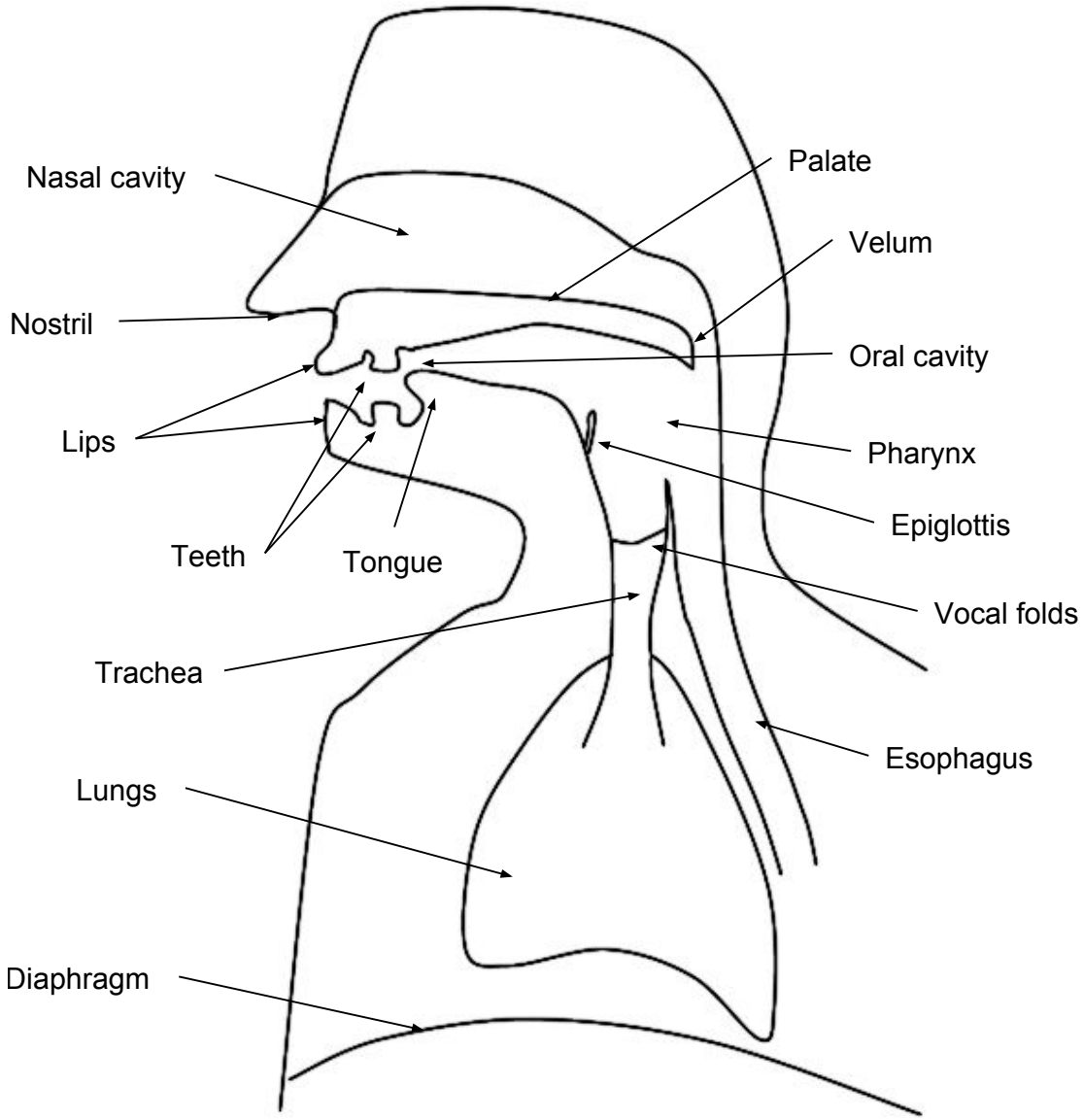



**Figure 3.2:** Cross-section diagram of the head and upper torso showing the locations of the various organs involved in the human speech production process.

organs are shown in a cross-sectional diagram of a human head and upper torso in Figure 3.2. The source of energy in human speech production comes from air expelled from the lungs using muscular force which generates an excitation signal when passing through the glottis (the opening between the vocal-folds within the larynx) through the process of phonation. This excitation waveform is subsequently modulated by the remainder of the vocal organs, termed the vocal-tract, to produce a sound pressure waveform emanating from the mouth. Different audio speech outputs are produced by engaging the various components of the vocal organs. The separation of the human speech production process into a source excitation signal and vocal-tract filter gives rise to the source-filter speech production model, which is discussed in greater detail in Section 3.3.

### 3.2.2 Excitation

Air-flow from the lungs can be used to produce two major types of excitation source. Voiced speech sounds are produced when the vocal-tract is excited with air-flow that is modulated by vibrations of the vocal folds. As air passes through the glottis the vocal folds begin to vibrate, thus modulating the flow of air to produce a periodic output. This process is known as phonation. The rate at which the vocal folds open and close determines the fundamental frequency, $f_0$, of the speech. Unvoiced sounds are produced through the excitation of the vocal-tract with air from the lungs flowing through the open vocal-folds and constrictions within the vocal-tract. The acoustic qualities of sounds produced from the excitation of turbulent air are more noise-like (aperiodic), with a broader spectrum than voiced excitation sounds. These two sources are not mutually exclusive, and when both occur together speech is produced with a more breathy quality. An additional source of excitation occurs when a build-up of pressure, due to the closing of a

part of the vocal-tract, is released. Such sounds occur for stop consonants.

### 3.2.3 Vocal-tract

The modulation of the excitation sound sources occurs through the process of acoustic resonance. The vocal-tract, beginning at the larynx and ending at the mouth and lips, is the primary resonant structure of the human speech production system. The main resonances that occur in the vocal-tract are known as formants, and are important in determining the phonetic properties of audio speech. Accordingly, the vocal-tract can be thought of as a filter with a series of resonances that modulate an excitation source. The first two formants, $F_1$ and $F_2$, are generally the most significant determinants of the phonetic properties of speech sounds, although for certain phonemes the higher-frequency formants are important [Holmes and Holmes, 2001]. Variation in $F_1$ and $F_2$ is based on the volume and shape of the pharyngeal and oral cavities, respectively. A larger pharyngeal cavity, as occurs when the tongue is raised, exhibits correlation with the first formant, whereas the second formant exhibits correlation with changes within the oral cavity [Cairns et al., 2010]. Furthermore, the nasal cavity can be incorporated into this system by disengaging the velum (soft-palate), which allows sound to radiate from the nostrils.

## 3.3 Speech production models

In the previous section, an overview was provided of the speech production system as occurs in humans, and how the process can be viewed as a source of excitation that is modulated by a filter producing an output signal. In this section, four popular speech production models are discussed with applications ranging

from speech coding to speech modification and synthesis. These four models include the sinusoidal model and a development known as the harmonic-plus-noise model, and then a typical source-filter model with a further implementation called STRAIGHT.

It is important to consider a variety of speech production models, as the one selected for this work will be used to produce audio speech reconstructions using spectral-envelope parameters obtained from the visual-to-audio domain mapping models, and excitation information from other methods. As it will be difficult to estimate the spectral-envelope accurately, due to the limitations of audiovisual correlation, it is important that the speech production model itself will not have a detrimental effect on the intelligibility of reconstructed utterances.

### 3.3.1 Sinusoidal model

The sinusoidal model of speech production was first proposed by McAulay and Quatieri [1986] and is based on the premise that each frame of speech, where the waveform is assumed to be stationary, can be represented by the summation of individual sinusoids each with varying amplitude, frequency, and phase. The model has application for low bit-rate speech coding [Marques et al., 1990], speech enhancement [Jensen and Hansen, 2001], and speech modification [George and Smith, 1997]. A generalised mathematical description of the sinusoidal model is given by,

$$s_i = \sum_{l=0}^{L-1} A_l \sin(\omega_l i + \phi_l), \tag{3.1}$$

where $\omega_l = 2\pi f_l$, and each frame of speech comprises $L$ sinusoids, each having frequency, $f_l$; amplitude, $A_l$; and phase, $\phi_l$. The set of $L$ sinusoidal components

can be extracted from the short-term Fourier transform of a speech frame using a peak-picking algorithm. For voiced frames of speech, the harmonics are usually sufficiently evident to be able to easily identify the peaks of each sinusoid, and determine the frequency and amplitude from the magnitude spectrum, and the phase from the phase spectrum. To allow the model to function during unvoiced speech frames, parameters are taken from sinusoids spaced 100 Hz apart.

During voiced speech frames there exists an approximate harmonic relationship between the sinusoidal components, where the sinusoids have frequencies that are integer multiples of the fundamental frequency. For example, with $f_0 = 120\,\text{Hz}$, sinusoids with frequencies of approximately 120 Hz, 240 Hz, ..., 3840 Hz, 3960 Hz will be present. Accordingly, Equation 3.1 can be simplified to give,

$$s_i = \sum_{l=0}^{L-1} A \sin(l\omega_0 i + \phi_l), \tag{3.2}$$

which exploits the approximate harmonic relationship of the sinusoids. For unvoiced frames as mentioned above, the fundamental frequency is chosen to be $f_0 = 100\,\text{Hz}$. The number of sinusoids generated per frame, $L$, is described by,

$$L = \left\lfloor \frac{f_s}{2f_0} \right\rfloor. \tag{3.3}$$

When synthesising speech using the per-frame sinusoid parameters, it is necessary that the amplitudes, frequencies, and phase are kept continuous across frame boundaries. To resolve these issues, frequency-matching algorithms are used in addition to phase-unwrapping and interpolation [McAulay and Quatieri, 1995].

### 3.3.2 Harmonic plus noise model

The harmonic plus noise model (HNM) is an extension of the sinusoidal model which aims to improve the naturalness and quality of the synthesised speech. The HNM can be implemented in one of two ways. Firstly, as in Laroche et al. [1993], where, as the sinusoidal model only synthesises sine waves at specific frequencies, with no energy present in between, the HNM attempts to improve the resultant audio quality by adding noise to "fill-in" the frequency gaps. Secondly, as in Stylianou [2001], where a frequency, $f_M$, is used as a cut-off separating the frequency spectrum in to two regions. The lower region of the spectrum describes the voiced speech, whereby only harmonics are used, and the upper region describes the unvoiced speech, which comprises purely noise with no harmonics. This separating of frequencies about a cut-off frequency is performed as the harmonic nature of the higher-frequency region is replaced with a more noise-like characteristic.

### 3.3.3 Source-filter model

The source-filter model [Rabiner and Schafer, 1978; Kleijn and Paliwal, 1995], is a notable model of speech production that separates the generation of audio speech into an excitation signal source and vocal-tract filter parameters. In practical applications, such as vocoders, the excitation signal takes the form of white-noise, with no controlling input for unvoiced speech, or pitch pulses, which are fundamental frequency dependent for voiced speech. A block diagram of the source-filter model is provided in Figure 3.3, showing sources of excitation, vocal-tract filter, and output audio signal.

A voicing decision (voiced/unvoiced) can be included to decide whether the excitation signal is to be generated using only either noise or pitch pulses. However, more realistic speech results from using a combination of both excitation sources

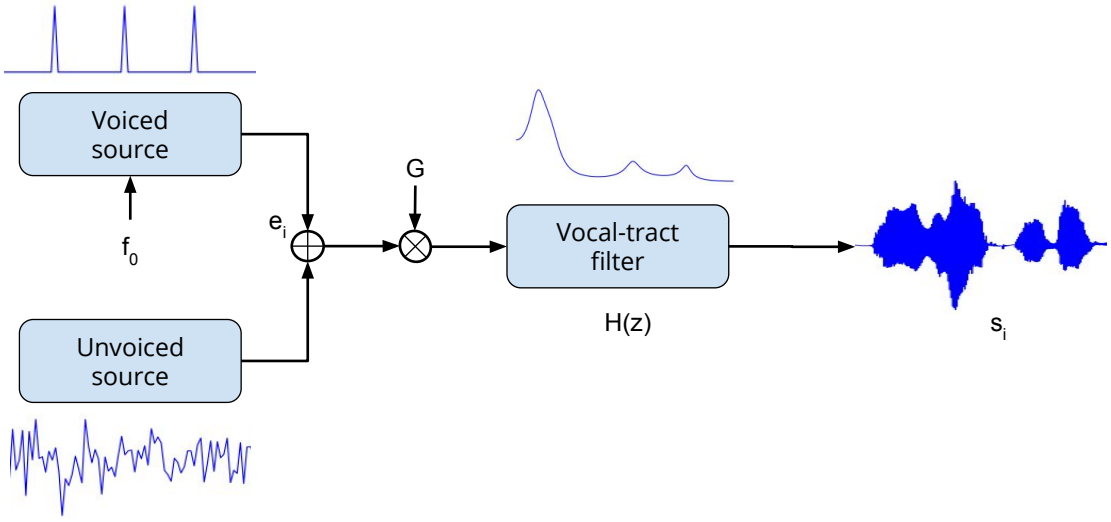



**Figure 3.3:** An overview of the source-filter model of speech production, showing voiced (glottal pulse train) and unvoiced (white noise) sources of excitation, gain term, vocal-tract filter response, and output audio signal.

as even during voiced speech the signal is not strictly periodic, with random fluctuations noticeable in the higher-frequency regions [Kawahara and Morise, 2011]. The periodicity of the pitch pulses is dictated by the fundamental frequency of the frame of speech.

A gain term is used to control the loudness of the window, and is determined from the energy within the window of speech. The spectral content of the excitation signal is then modulated by the filter, which takes as input a number of filter coefficients derived from the vocal-tract. A speech signal can be obtained through application of,

$$s_i = \sum_{p=1}^{P} a_p s_{i-p} + G \cdot e_i, \tag{3.4}$$

where $a_p$ is the $p^{th}$ coefficient of the all-pole vocal-tract filter with order $P$, $G$ is a gain term, and $e_i$ is an excitation signal. A voicing decision, necessary for selecting

the excitation source (white-noise or pitch-pulses), and fundamental frequency, used for setting the period of pitch-pulses during voiced speech, can be obtained using pitch detection algorithms such as PEFAC [Gonzalez and Brookes, 2014] or YIN [De Cheveigné and Kawahara, 2002]. The filter coefficients and gain term can be determined through signal processing techniques such as linear predictive coding, which is discussed further in Section 3.4.

### 3.3.4 STRAIGHT

The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) toolkit, proposed originally by Kawahara et al. [1999] and receiving major updates in Kawahara et al. [2008], is an implementation of the source-filter model that separates speech into its spectral-envelope and source excitation components. The model was developed to allow for flexible manipulation of parameters to produce high-quality speech modifications. The speech model has received considerable attention in the areas of text-to-speech synthesis, where it has been used for a number of HMM-based statistical speech synthesis systems [Yamagishi et al., 2007, 2009; Heiga et al., 2007], and for voice conversion systems [Toda et al., 2001; Ohtani et al., 2006].

To synthesise a time-domain speech signal, STRAIGHT requires a set of three parameters:

1. a fundamental frequency contour – $f_{0_i}$,
2. a measure of aperiodicity – $A(f, i)$,
3. and a spectral-envelope surface – $X(f, i)$,

where $i$ and $f$ represent the frame index and frequency bin respectively. As an

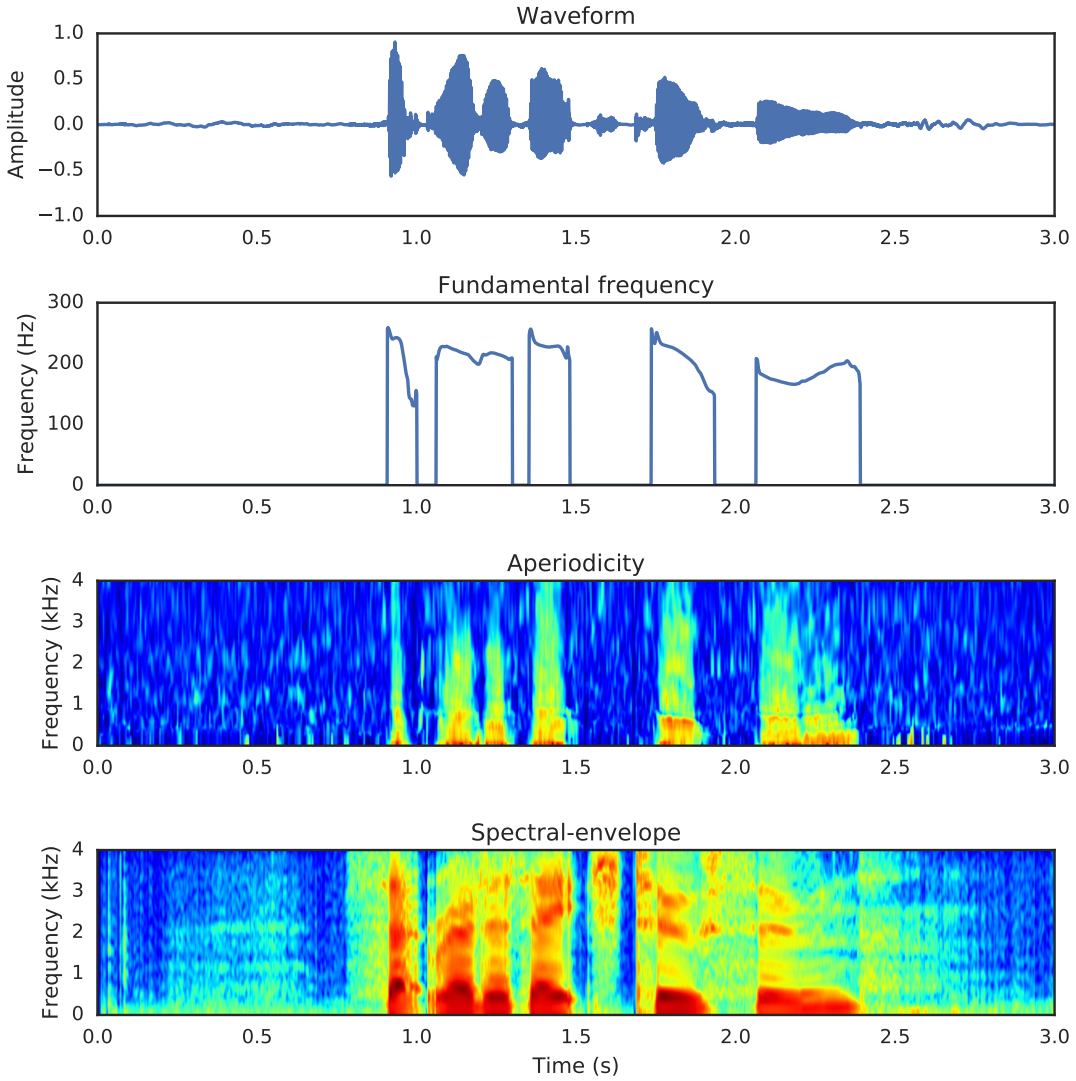



**Figure 3.4:** Original waveform of the utterance "set red at H 2 soon" and parameters extracted using STRAIGHT of fundamental frequency, aperiodicity, and spectral-envelope.

example, a set of parameters extracted from a speech utterance are shown in Figure 3.4. The aperiodicity surface provides a measure of how periodic (or aperiodic) the frequency components are in the reconstructed speech signal. During speech production, aperiodic sounds are more noise-like (having more aperiodic components) and are produced by means of aspiration, frication, and transient bursts due to constrictions in the vocal-tract [Deshmukh et al., 2005]. In comparison, periodic sounds produced during voiced speech are created by the vibration of the vocal folds. Additionally, the filtering of these sound sources by the vocal-tract further affects the aperiodicity of the frequency components. The measure of aperiodicity allows for both periodic and aperiodic components to be combined, as is the case for voiced fricatives, to give more natural sounding speech.

To allow for high-quality speech reproductions, STRAIGHT performs successive refinements of the source and spectral parameters. A pitch-adaptive smoothing algorithm is applied to the spectral-envelope surface, $X(f, i)$, to remove interference caused by periodic components in the frequency and time domains. Furthermore, as speech reconstructed using simple channel vocoders can exhibit a buzzy-quality due to the characteristics of the input glottal pulse-train excitation source, an all-pass filter is used to better control the temporal structure of the pulse-train [Kawahara, 1997].

The STRAIGHT toolkit provides both a source-filter and a sinusoidal implementation for resynthesising a speech signal. The source-filter approach is used in this work, which requires vocal-tract filter and source excitation information. The vocal-tract filter impulse response, $h_i(t)$, can be obtained using,

$$h_i(t) = \mathcal{F}^{-1}\big[H_i(f)\Phi_i(f)\big], \tag{3.5}$$

where $\mathcal{F}^{-1}$ signifies the inverse Fourier transform, $H_i(f)$ is the Fourier transform

of the minimum phase impulse response derived from the spectral-envelope surface [Kato, 2017], and $\Phi_i(f)$ is the all-pass filter. The minimum phase impulse response is desired as specific types of phase characteristics can have an adverse affect on the resultant reconstructed speech [Kawahara et al., 1999].

The excitation signal, $e_i(t)$, is obtained by combining both periodic and aperiodic sources of excitation as follows,

$$e_i(t) = \frac{1}{\sqrt{f_{0_i}}}\delta(n \bmod \frac{fs}{f_{0_i}}) + \mathcal{F}^{-1}\big[A_i(f)|N(f)|\big], \tag{3.6}$$

where the delta function is described by,

$$\delta(x) = \begin{cases} 1 & \text{if } \lfloor x \rfloor = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

$|N(f)|$ is the magnitude spectrum of random white noise, and $fs$ is the sampling frequency. The first term in Equation 3.6 gives the periodic source of excitation by using the delta function to produce pitch-pulses at the fundamental frequency, $f_{0_i}$. The second term gives the aperiodic excitation source by taking the inverse Fourier transform of the magnitude spectrum of random white noise multiplied by the aperiodicity surface.

A frame of speech, $y_i(t)$, can be reconstructed by convolving the excitation source, $e_i(t)$, with the vocal-tract filter response, $h_i(t)$, as follows,

$$y_i(t) = h_i(t) * e_i(t), \tag{3.8}$$

where $*$ denotes the convolution operator. To produce the entire utterance, overlap-and-add is applied to the sequence of frames to give the final output speech signal.

Given the successful application of STRAIGHT in a number of areas of speech processing, and its ability to reconstruct high-quality speech signals, the speech production model has been chosen for generating intelligible audio speech utterances in this thesis.

## 3.4 Audio features

There are two primary considerations to take into account when selecting the audio feature representations of the spectral-envelope information for use in the visual-to-audio domain mapping models. Firstly, that there exists sufficient correlation between the audio and visual features, so that they can be estimated with high accuracy using mapping models. Secondly, that a suitable spectral-envelope surface can be reconstructed from the audio feature coefficients, for use within the STRAIGHT speech production model.

Previous studies have shown there to exist good correlation between visual information and audio features such as mel-frequency cepstral coefficients (MFCCs) [Almajai et al., 2006], also commonly used as the front-end for ASR systems, and the line spectral pairs (LSPs) representation of linear predictive coding (LPC) coefficients [Yehia et al., 1998; Barker and Berthommier, 1999], commonly used in speech coding tasks. In this section, LPC coefficients and mel-filterbank channel amplitudes are discussed, with equations detailing how spectral-envelope reconstructions are obtained from each type of feature.

### 3.4.1 Linear predictive coding coefficients

Linear predictive coding (LPC) is a common analysis technique for estimating vocal-tract filter coefficients, and has application in tasks requiring codifying of

speech signals, such as in the CELP [Schroeder and Atal, 1985] and SILK [Vos et al., 2010] speech codecs. The principal behind linear prediction is that future values of a discrete-time signal, produced by a slowly varying linear filtering process, can be estimated as a linear combination of previous values [O'Shaughnessy, 1988]. LPC analysis is able to produce compact and precise representations of the magnitude spectrum for short signals, where it is assumed that the signal is briefly stationary, which for speech relates to configurations of the vocal-tract. For most LPC analysis, it is satisfactory to assume that the filter is an all-pole model, where the signal spectrum can be adequately modelled by resonances indicating the spectral peaks.

An all-pole spectral shaping filter, $H(z)$, with order $P$, can be described by:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{3.9}$$

where $a_k$ are the filter coefficients, and $z^{-k}$ is a delay of $k$ samples. To derive a set of coefficients, $a_k$, for parametrising the all-pole model, $H(z)$, two least-squares methods can be applied, *autocorrelation* and *covariance*. To suitably model the formants in a frame of speech, two or more poles are required per resonance, where, in practice, for an 8 kHz sampling frequency, ten poles are typically adequate [O'Shaughnessy, 1988].

To obtain a spectral-envelope from a set of LPC coefficients, $a_k$, the frequency response of the filter is evaluated at equally spaced frequencies up to the bandwidth, through application of,

$$X^{\mathrm{LPC}}(f, i) = 10 \log_{10} \left| H_i(e^{j2\pi f}) \right|^2, \tag{3.10}$$

where to reproduce the power spectrum, assuming the set of linear predictor co-

efficient were obtained from the magnitude spectrum, it is necessary to perform additional logarithmic and power operations. In Chapter 5, the linear predictor order is evaluated to find an optimum size.

### 3.4.2 Filterbank channel amplitudes

Filterbank representations of speech signals encode the high-resolution information of the frequency domain as a low-dimensional feature vector, where the coefficients are outputs taken from a bank of bandpass filters with the objective of retaining the most perceptually important information. The bandwidths and centre frequencies of the filterbank channels are typically chosen to increase with frequency, and are motivated by auditory filter models of the frequency resolving abilities of the *cochlea*. The spacing and bandwidth of the channels is commonly chosen to conform to perceptual scales such as the Bark or mel scales. Filterbank channel amplitudes, with a mel spacing, for speech recognition applications were first advocated by Davis and Mermelstein [1980]. Cepstral analysis can be applied to Mel-spaced filterbank amplitudes yielding the popular Mel-frequency cepstral coefficient (MFCC) audio feature, frequently used in ASR applications.

To obtain a set of Mel-filterbank channel amplitudes, $\mathbf{a}_i^{\mathrm{MEL}}$, the ETSI Aurora standard [ETSI, 2002] is followed. First, a bank of triangular bandpass filters is applied to the short-term magnitude spectrum of a frame of speech. The spacing and bandwidths of the channel conforms to the mel scale, with a channel number of 23 typically used for speech processing applications having a sampling frequency of 8 kHz. The frequency energies within each band are summed to give a single output for that particular channel. The logarithmic transform of the channel amplitudes is then performed, motivated by the compressive non-linearity of the *basilar membrane*, whereby a large range of input sound pressure levels are

compressed into a smaller range [Holmes and Holmes, 2001].

A spectral-envelope representation can be reconstructed from a set of filterbank channel amplitudes, $\mathbf{a}_i^{\text{MEL}}$, through application of,

$$X^{\text{MEL}}(f,i) = \text{interp}(\sqrt{e^{\mathbf{a}_i^{\text{MEL}}}}), \tag{3.11}$$

where linear interpolation is applied to the transformed channel amplitudes at the mel-spaced frequencies, to convert from the non-linear frequency spacing of the filterbanks to a linear spacing, resulting in an spectral-envelope covering the frequency range of 0 to 4 kHz. Experiments are conducted in Chapter 5 to determine the optimum number of channels required for this work.

## 3.5 Visual features

Raw visual information is collected in the form of video recordings of a speaker's face, where, for even low-resolution video footage, there will be a large number of pixel intensities. Furthermore, unless the video is capturing only the mouth of a speaker, the majority of pixels within each frame will be redundant. Accordingly, for encoding visual speech information, it is not necessary to use all of the information within each frame, and, therefore, this high-dimensional data can be transformed into a considerably lower-dimension without any loss of important information. Numerous feature types exist for representing the visual speech information of a speaker, and can be broadly classified into four categories [Zhou et al., 2014]:

1. Motion-based – where the observed motion of the mouth over time is encoded,

2. Geometric-based – where combinations of the height, width, area, and perimeter of the mouth are extracted,

3. Image-based – where pixel intensities located about the mouth are either used directly or after application of an image transform,

4. Model-based – where parameters are derived from a compact statistical model describing the shape and appearance of the mouth.

Two visual feature representations that have been successfully applied in various areas of audiovisual speech processing are the image-based two-dimensional discrete cosine transform (2D-DCT), and the model-based active appearance model (AAM). Both types of feature have been explored for AVSR [Potamianos et al., 2003] and automatic lip reading [Lan et al., 2009].

### 3.5.1 Two-dimensional discrete cosine transform

The discrete cosine transform (DCT) has application in a number of areas of signal processing and image coding where signal compression is desired [Rao and Yip, 2014]. The idea behind the DCT is that a signal can be expressed as a weighted summation of a number of cosine functions each with a different frequency, where the majority of the signal information is concentrated within the low-frequency coefficients of the transformed signal. When used in signal processing applications, such as for lossy compression in the MP3 audio codec, the high-frequency coefficients can be discarded as they are deemed to be perceptually unimportant. The two-dimensional discrete cosine transform (2D-DCT) has application for lossy compression of images, and is implemented in the JPEG imaging coding standard.

For use in visual speech applications, 2D-DCT features are extracted from a matrix of pixel intensities, $\mathbf{P}$, centred on the mouth of a speaker, yielding a coefficient

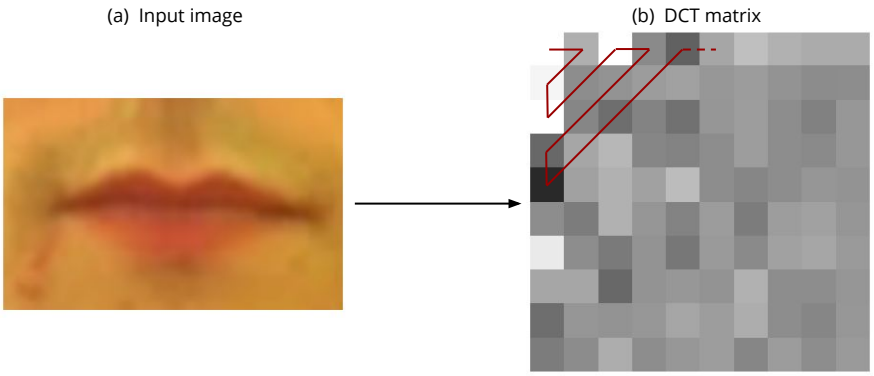



**Figure 3.5:** An input image of a speaker's mouth is shown in (a), and the corresponding top left of the DCT matrix is shown in (b). The application of zigzag scanning is shown by the red line in (b).

matrix, $\mathbf{C}$, through application of,

$$\mathbf{C}_{v,u} = 4 \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} \mathbf{P}_{y,x} cos\Big[\frac{\pi(2y+1)v}{2M}\Big] cos\Big[\frac{\pi(2x+1)u}{2N}\Big], \quad 0 \leq v \leq M-1, 0 \leq u \leq N-1, \tag{3.12}$$

where $M$ is the number of rows and $N$ is the number of columns of matrices $\mathbf{P}$ and $\mathbf{C}$. After application of Equation 3.12, the output coefficient matrix $\mathbf{C}$ has the same dimensionality as the input matrix $\mathbf{P}$, where the low-frequency information is concentrated in the upper-left corner, and the high-frequency information is concentrated in the lower-right corner. The application of the 2D-DCT transform to an example mouth image is shown in Figure 3.5. To convert matrix $\mathbf{C}$ to a vector, zigzag-scanning is applied beginning at the low-frequency region [Sayood and Fow, 2000]. This yields a DCT coefficient vector,

$$\mathbf{c} = [c_{0,0}, c_{0,1}, c_{1,0}, ..., , c_{M-1,N}, c_{M,N-1}, c_{M,N}], \tag{3.13}$$

which can subsequently be truncated to produce an output $J$-dimensional visual feature vector, $\mathbf{v}_i^{\text{2D-DCT}}$.

### 3.5.2 Active appearance model

Proposed by Cootes et al. [2001], an active appearance model (AAM) represents the shape and appearance of an object, and has application for tracking objects described by the model in an new image, such as for locating faces. An AAM is constructed from a set of training images and associated hand-labelled landmark data describing the objects of interest within each image, where the training images are typically chosen so as to cover the range of variations exhibited by the object. Principal component analysis (PCA) is applied to shape and appearance parameters determined from the training data to produce a model. In audiovisual speech processing applications, an AAM can be used to model the shape and appearance of a speaker's mouth, an example of which can be seen in Figure 3.6, from which visual features can then be extracted.

**Figure 3.6:** Shape, shown as the red line on the inner and outer lip contours, and appearance information of a speaker's mouth.

The shape parameter, $\mathbf{s}$, is the concatenated coordinates of $n$ vertices detailing the contours of the inner and outer lips, $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^\intercal$, and can be

described by,

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q}_s \mathbf{b}_s, \tag{3.14}$$

where $\bar{\mathbf{s}}$ is the mean shape of the model, matrix $\mathbf{Q}_s$ describes the modes of shape variation derived from the training data, and $\mathbf{b}_s$ defines the contribution of each mode. The columns in $\mathbf{Q}_s$ are the leading eigenvectors of the covariance matrix defining these modes of variation, or principal components. For a shape model of a mouth, these principal components capture variation such as rounding of the lips, and opening and closing of the mouth [Newman et al., 2010].

The appearance parameter, $\mathbf{a}$, is defined by the pixel intensities located within the mean shape, $\bar{\mathbf{s}}$, after the shape of the input image is normalised, and can be described by,

$$\mathbf{a} = \bar{\mathbf{a}} + \mathbf{Q}_a \mathbf{b}_a, \tag{3.15}$$

where $\bar{\mathbf{a}}$ is the mean appearance image of the model, matrix $\mathbf{Q}_a$ describes the modes of appearance variation derived from the training data, and $\mathbf{b}_a$ defines the contribution of each of the variation modes. The principal components of the appearance model capture variations in texture.

To extract visual features of a speaker's mouth from an input visual frame using an AAM, the model is first applied to determine a set of tracked landmarks located about the inner and outer lips of the speaker. These landmarks are then processed using Equation 3.14 to derive the shape parameters, and after warping the image to the mean shape, the appearance parameters are extracted using Equation 3.15. A final application of PCA is performed to the concatenated shape and appearance parameters to produce an output AAM visual feature vector, $\mathbf{v}_i^{\mathrm{AAM}}$.

## 3.6 Summary

The aim of this chapter has been to identify and select an appropriate speech production model, and suitable audio and visual speech feature representations, for developing the work in this thesis. To begin, an overview of the human speech production process is given, discussing the anatomy of the vocal organs and their use in producing audio speech signals. The production of speech can be viewed as a source-filter model, whereby an excitation source is modulated by a time-varying filtering process. The excitation source takes the form of pitch-pulses produced when the vocal-folds vibrate for voiced sounds, and of turbulent air due to constrictions in the vocal-tract for unvoiced sounds. The vocal-tract filter modulates the excitation by introducing acoustic resonances, that are dependent on the configuration of the articulators, and can be seen in the frequency domain as peaks, otherwise known as formants. This source-filter model view of speech production is exploited in speech production models.

Following this, four speech production models are reviewed, with STRAIGHT being identified as the best choice for this work, due to its ability to produce high-quality speech reconstructions and as it has received significant attention for numerous speech processing tasks in the literature. The necessary parameters for the model include excitation information in the form of a fundamental frequency contour and aperiodicity surface, and vocal-tract configuration in the form of a spectral-envelope surface. The application of all-pass filtering to the glottal pulse-train during voiced speech yields a more natural temporal structure, with the pitch-adaptive smoothing applied to the spectral-envelope surface removing periodic interferences that affect the quality of other speech models.

As this work relies on the ability to map between the visual and audio domains, two audio and two visual feature representations are considered, with reference to

the correlations that exists between the modalities and, additionally, the features in question. Linear predictive coding coefficients are able to model the spectral-envelope surface and formants therein providing an adequate order is chosen. Filterbanks are motivated by the auditory-filter model of the cochlea, where using a mel-scale yields greater resolution at the lower frequencies, and where applying the log transform to the channel amplitudes reproduces the non-linear compressibility of the basilar membrane. Unlike in the audio domain, where there is an obvious choice for audio speech representations, for audiovisual and visual-only tasks there is little consensus on the best visual features to use. Two features that have shown good performance for AVSR and lip-reading tasks include the image-based 2D-DCT features, and the model-based AAM features, where the AAM features are typically shown to perform best for representing visual speech information.

In the next chapter, the work progresses onto the excitation information as required by STRAIGHT. A small number of audio-only experiments are conducted to motivate excitation choices, and to further explore the two spectral-envelope features described here. In Chapter 5, combinations of the audio and visual features, using two mapping models, are explored, with decisions made as to which configurations to focus on.

# Chapter 4

# Excitation

## 4.1 Introduction

In this chapter, an overview is provided of various methods for producing excitation information, fundamental frequency contour and aperiodicity surface, as is required by the STRAIGHT speech production model. The fundamental frequency contour describes the first harmonic for voiced sections of a speech signal, and is the frequency at which the vocal folds open and close. During sections of voiced speech, the $f_0$ contour will fluctuate around the average fundamental frequency of a speaker, whereas during unvoiced sections, no fundamental frequency exists as the vocal folds do not vibrate. In practice, the $f_0$ contour during unvoiced (and non-speech) sections is zero. The aperiodicity surface describes the relative aperiodicity of signals over the frequency domain. Generally, during voiced sections there is less aperiodicity as speech is produced from the vibration of the vocal folds, whereas during unvoiced sections, the aperiodicity is greater as speech is produced from turbulent air, resulting in greater noise-like characteristics. Obtaining spectral-envelope information will be discussed in Chapters 5, 6, and 7.

It is not possible to derive fundamental frequency and aperiodicity values from visual speech as the articulators which can be "seen" provide little, if any, information with regards to these. Perhaps the only information that can be extracted is a voicing decision (non-speech, unvoiced, or voiced) due to the visual realisations of voiced and unvoiced phonemes, although there is likely to be a large number of confusions. Accordingly, in this chapter, three artificial methods are proposed and evaluated for providing fundamental frequency contour values. These methods are inspired by work from Miller et al. [2010] on the intelligibility of speech signals with various $f_0$ modifications. The generation of an aperiodicity surface is then explored using two methods. The first method investigates voicing classification of visual speech using neural networks and convolutional neural networks, where non-speech, unvoiced, and voiced aperiodicity vectors are selected based on estimated class labels. The second model of aperiodicity estimation uses vector quantisation techniques, where a codebook is produced from joint spectral-envelope and aperiodicity vectors, where aperiodicity estimates can be output given input spectral-envelope vectors. Both methods are evaluated to determine the most suitable choice.

The remainder of this chapter is organised as follows. In Section 4.2, an overview of the proposed artificial fundamental frequency contour methods is given. Visual voicing classification for producing an aperiodicity surface using neural networks, including using convolutional neural networks for visual feature extraction, is presented in Section 4.3. Details of the second aperiodicity estimation model using vector quantisation techniques are given in Section 4.4. Finally, the various excitation information methods and models are evaluated in Section 4.5, with conclusions drawn in Section 4.6.

## 4.2 Artificial-$f_0$ methods

In the STRAIGHT speech production model, the source excitation information is split into two separate parameters: fundamental frequency and aperiodicity. To produce the necessary fundamental frequency contour for an utterance, three artificial-$f_0$ methods are presented:

1. monotone,
2. time-varying,
3. unvoiced.

To provide values for the monotone and time-varying methods, an analysis is performed of the fundamental frequency contours of a speaker. The $f_0$ contour is extracted from each training utterance of a speaker where the non-voiced frames are ignored in the analysis. The short-time fast Fourier transform (STFT) of the signal is taken to provide approximate estimates of the parameters used for the time-varying method, although a certain amount of trial-and-error is required to find suitable values.

### 4.2.1 Monotone

The monotone method imitates monotonic speech by using a constant fundamental frequency value for all frames of the utterance. To derive a constant value, $f_0$ contours are extracted from training utterances of a speaker, with the mean taken of the $f_0$ values from voiced sections of speech, giving

$$f_{0_i} = \mu_{f_0}. \tag{4.1}$$

From the fundamental frequency analysis, it was determined that the appropriate mean fundamental frequency values, $\mu_{f_0}$, were 100 Hz and 207 Hz for the male and female speakers used in this work, respectively.

### 4.2.2 Time-varying

The time-varying method is motivated by experiments conducted by Miller et al. [2010] where one of the $f_0$ modifications is to use a slowly-oscillating sinusoid. The method modulates the monotone $f_0$ contour from Equation 4.1 using a 0.25 Hz cosine wave with an amplitude that gives a frequency change, $\Delta_{f_0}$, of $\pm$17.5 Hz for the male speaker, and $\pm$28 Hz for the female speaker. The delta values for each speaker are taken as the standard deviations of the mean fundamental frequency analysis described previously for the monotone method, and the oscillation frequency was determined using informal listening tests. An $f_0$ contour can be produced using

$$f_{0_i} = \mu_{f_0} + \Delta_{f_0} \cos\Big[\frac{2\pi i}{400} + \phi_r\Big], \tag{4.2}$$

where $\phi_r$ is a random phase offset in the range $-\pi$ to $\pi$. The phase offset is included to ensure that the beginning of each utterance does not start with an immediate decrease in frequency value due to the standard output of the cosine function. This is performed in an attempt to produce more natural $f_0$ contours.

### 4.2.3 Unvoiced

The unvoiced method uses fundamental frequency contour values of zero. Although in reality no values exist for $f_0$ during sections of unvoiced speech, for implementation purposes a value of zero is used. Speech reconstructed using an unvoiced contour yields utterances where the excitation source is white noise.

### 4.2.4 Method analysis

To provide better intuition for the three artificial-$f_0$ methods, a brief analysis is provided including a visual description of the contours produced by the monotone and time-varying methods in comparison to the original, and spectrograms of utterances reconstructed using the three methods. Figure 4.1 shows a waveform of an utterance from the female speaker, and the corresponding fundamental frequency contours for the original ground-truth, and the monotone and time-varying methods. For each of the artificial-$f_0$ methods, no consideration is given as to whether frames are voiced, unvoiced, or non-speech. That is, the values generated by each of the methods is used for all frames in the utterance. For the monotone method, the single value can be seen throughout the utterance, whereas for the time-varying method the slow-oscillation of the contour is apparent.

Narrowband spectrograms of the same utterance reconstructed using the original and three artificial-$f_0$ methods are shown in Figure 4.2. For the monotone and time-varying methods the harmonics can be seen, where they are constantly spaced for the monotone method, and slowly-oscillating for the time-varying method. In certain sections of the signal the time-varying method appears to follow quite faithfully the general structure of the original contour, although it is also visible that the fine-grained contour changes are missing. For the unvoiced method, there is no apparent harmonic structure, which is expected as the signal is excited with white noise.

From informal listening tests, the sound of utterances produced by the monotone method exhibit a slight buzzy quality with a robotic character, although remain relatively faithful to the original utterances. The oscillating $f_0$ contour of the time-varying method is quite noticeable and unnatural sounding, especially in sections where it does not mimic the original contour. Utterances produced using

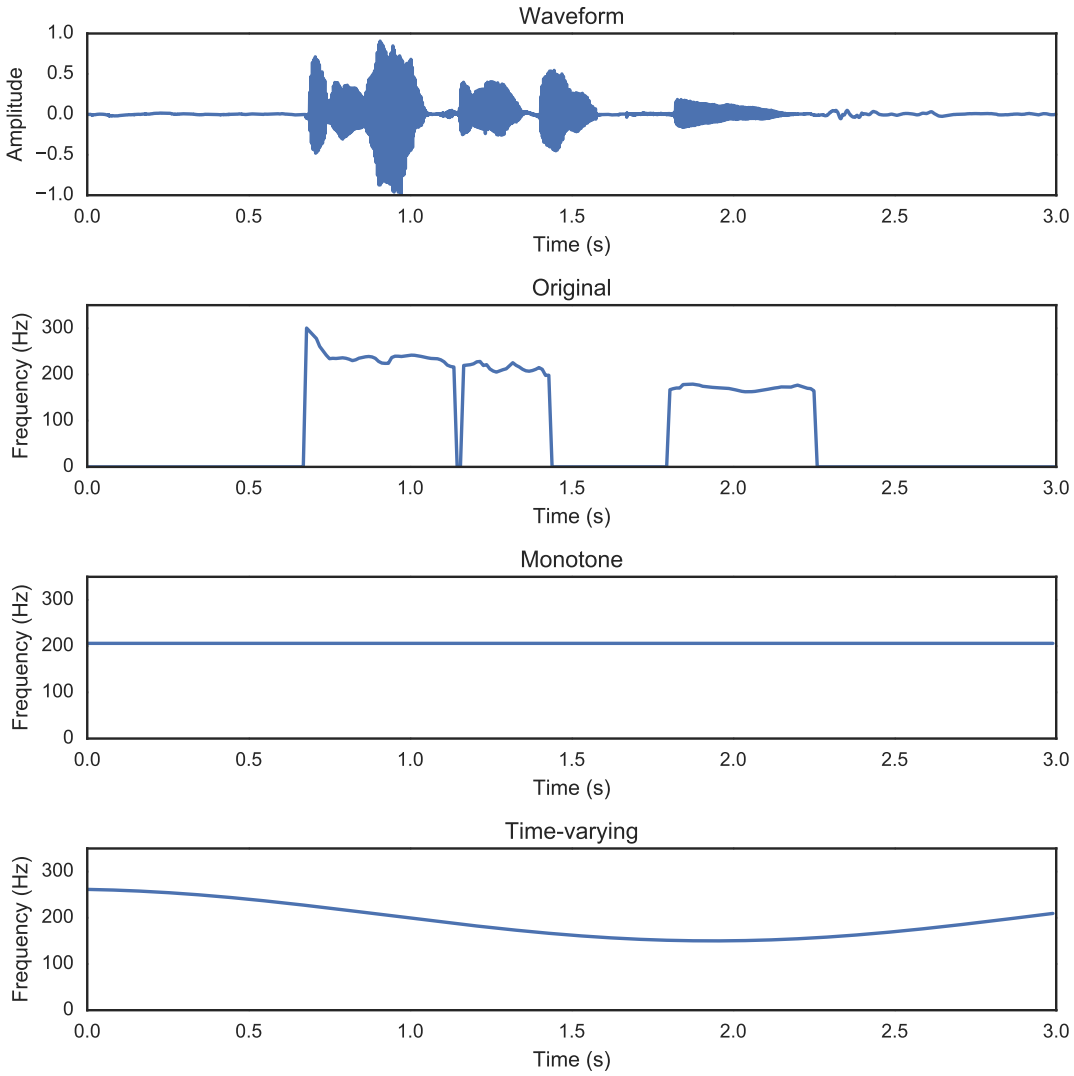



**Figure 4.1:** Comparison of the utterance "bin white in F 8 soon" spoken by the female speaker, showing the original waveform and ground-truth $f_0$ contour, and contours produced by the monotone and time-varying artificial-$f_0$ methods.

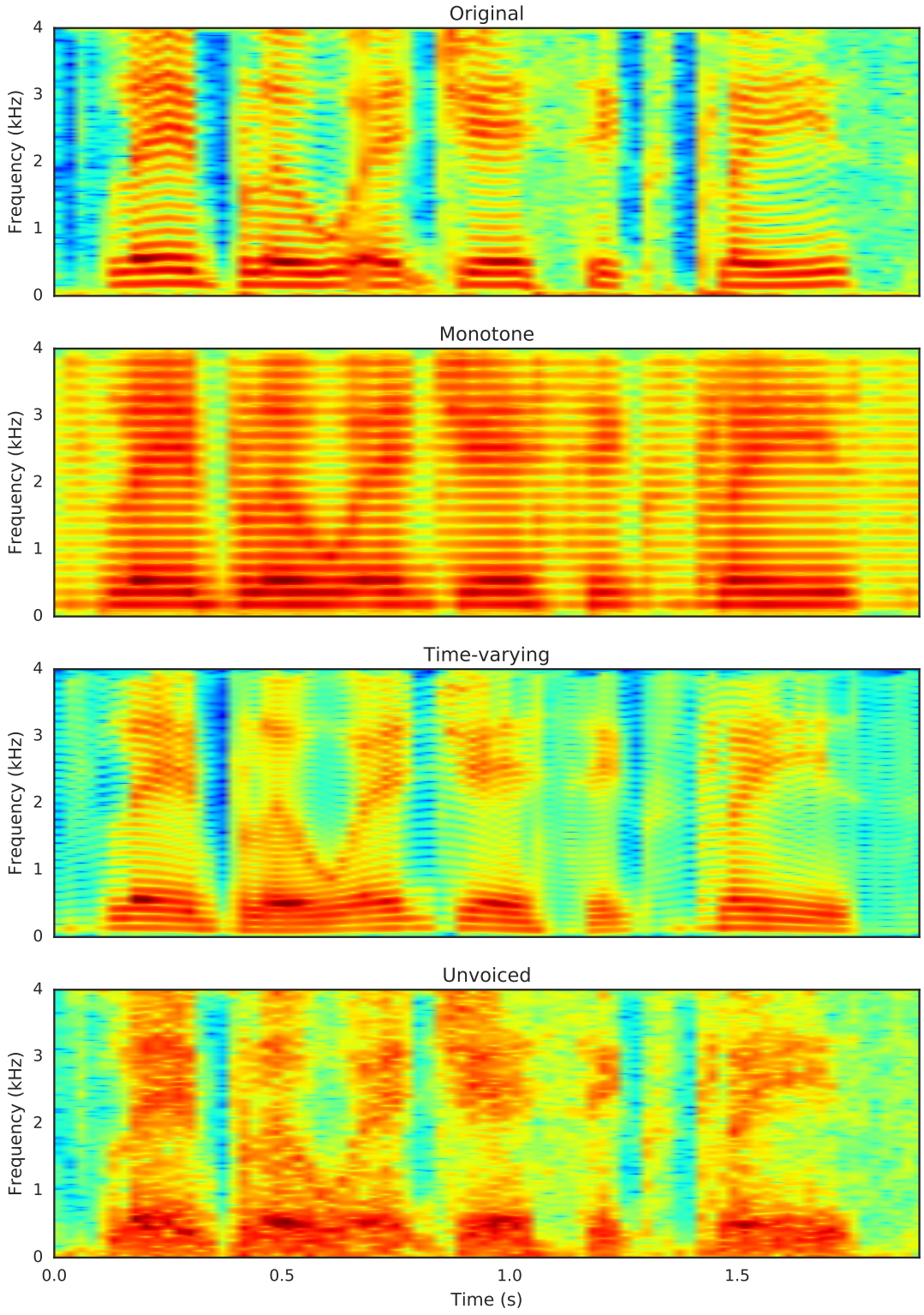



**Figure 4.2:** Narrowband spectrograms for the utterance "bin white in F 8 soon" spoken by the female speaker, with reconstructions using the original and three artificial-$f_0$ fundamental frequency contour methods.

the unvoiced method have a harsh, raspy character that is emphasised by sections of the speech signal with high energy.

Reconstructed utterances using contours generated from the three proposed artificial-$f_0$ methods are evaluated using subjective listening tests briefly in Section 4.5, and in more detail in Chapter 5. Having developed methods for producing the required fundamental frequency information, the next two sections detail approaches to estimating aperiodicity information.

## 4.3 Aperiodicity estimation using visual voicing classification

In this section, the first approach to aperiodicity estimation is described, where voicing classification models applied to input visual speech features are used to output voicing class labels. The predicted class labels are then used to select either non-speech, unvoiced, or voiced aperiodicity vectors. Two neural network architectures are explored for performing voicing classification. The first method estimates class labels using a standard single hidden-layer neural network from input 2D-DCT visual features, and the second uses convolutional neural networks (CNN) applied to raw pixels located about the mouth of the speaker for performing visual feature extraction.

Voicing classification is the challenge of classifying speech frames (either audio, visual, or audiovisual) as being either non-speech, unvoiced, or voiced. In this work, the aim is to learn a function, $f$, to estimate the voicing class, $\hat{c}_i^{\text{VC}}$, of the input visual speech feature vector, $\mathbf{v}_i$, described by

$$\hat{c}_i^{\text{VC}} = f(\mathbf{v}_i), \tag{4.3}$$

where $\hat{c}_i^{\text{VC}} \in \{\text{ns, u, v}\}$ for voicing classification. Grouping the unvoiced and voiced class labels together allows for a voicing classification system to function for voice activity detection, where the problem is of determining speech and non-speech frames.

In the remainder of this section, the standard and convolutional neural network architectures are described, including techniques for performing regularisation to ensure the networks do not overfit on the training data. The generation of aperiodicity surfaces is then explained using these voicing classification models. These experiments also serve as an investigation into using CNNs for visual feature extraction, where motivations for their use are provided.

### 4.3.1 Neural network

Neural networks are a group of learning algorithms loosely based on the biological operations of neurons in the brain. Inputs are fed through a series of layers comprised of individual units (neurons), where a non-linear activation function is then applied to the output of certain layers. An example fully-connected neural network, where the units in layer $m$ are connected to all of those in layer $m - 1$, is shown in Figure 4.3a. The hidden layers perform feature extraction by learning non-linear combinations of the inputs, where individually the features may not be particularly descriptive [Murphy, 2012]. Care must be taken when training neural networks as they are prone to over-fitting on the training set if there is a lack of training material. In this section, the use of neural networks is described for performing voicing classification from input 2D-DCT visual speech features.

To estimate the voicing class of a frame given input visual features, a feed-forward neural network model can be used for function $f$ in Equation 4.3. The function $f$ is comprised of a single hidden layer between the input and output

layers, with the model weight parameters derived from a set of training data using the backpropagation of errors algorithm.

The output of the hidden layer, $\mathbf{h}$, is a function of the input visual parameters, $\mathbf{v}$, and the weight connections between the two layers, $\mathbf{W}_{hv}$. A bias term, $\mathbf{b}_v$, is included to provide each neuron in the input layer with a constant output, performing a similar role to the intercept in standard linear regression. In practice, however, the bias terms are usually incorporated into the weight parameter matrix. The output from the hidden layer can be obtained from

$$\mathbf{h} = \sigma\left(\mathbf{W}_{hv}^{\intercal}\mathbf{v} + \mathbf{b}_h\right), \tag{4.4}$$

where $\sigma$ is a non-linear and differentiable activation function such as the sigmoid (logistic) or tanh functions, or the rectified linear unit (ReLU). The ReLU function is a non-saturating activation function proposed by Nair and Hinton [2010], and is calculated as $\sigma(x) = \max(0, x)$. Conversely, the tanh and sigmoid functions both saturate given large input values. One benefit of building neural networks using the ReLU activation function is that training concludes several times faster over sigmoid activations [Krizhevsky et al., 2012]. It is important that a non-linearity is applied after the weight multiplications as otherwise the network will learn functions that are linear combinations of the inputs. Furthermore, the activation function is required to be differentiable as the gradient descent method is used for training of the weight parameters.

To obtain a voicing class estimate, $\hat{c}_i^{\mathrm{VC}}$, from a feed-forward neural network with a single hidden layer architecture, a visual feature vector, $\mathbf{v}_i$, is presented as input to

$$\hat{\mathbf{z}}_i = \mathrm{softmax}\big[\mathbf{W}_{oh}\sigma\left(\mathbf{W}_{hv}\mathbf{v}_i\right)\big], \tag{4.5}$$

where $\mathbf{W}_{hv}$ is the weight connection matrix between the input layer and hidden layer, and $\mathbf{W}_{oh}$ is the weight connection matrix from the hidden to output layers applied to the result of the hidden layer after application of the non-linear activation function, $\sigma$. The softmax function is applied to the outputs of the final layer to give a set of posterior probabilities for the output classes conditioned on the input visual features, i.e. the *a posterior* probability $p(c_j|\mathbf{v})$ is the probability of codebook entry $c_j$ given the input visual feature vector $\mathbf{v}$.

Given a $K$-dimensional real-valued vector, $\mathbf{x}$, the softmax function can be applied to obtain $K$ class probabilities, through application of

$$\mathbf{z}_j = \text{softmax}(\mathbf{x}_j) = \frac{\mathrm{e}^{x_j}}{\sum_{k=1}^{K} \mathrm{e}^{x_k}}, \tag{4.6}$$

where each element in $\mathbf{z}$ lies within the range $(0, 1)$, and all coefficients sum to a total of one. To obtain the estimated class label, $\hat{c}_i^{\text{VC}}$, with the greatest probability, the arg max can be taken over the output class probabilities, $\hat{\mathbf{z}}$, using

$$\hat{c}_i^{\text{VC}} = \arg\max_z \hat{\mathbf{z}}_i. \tag{4.7}$$

To derive the required weight parameters for each of the layer connections, the backpropagation of errors algorithm, used in conjunction with gradient descent optimisation, is applied to minimise the categorical cross-entropy between the output of the final softmax layer and correct class labels.

Cross-entropy, from the field of information theory, is used as the basis of the cost function, which provides a measure of similarity between two probability distributions. More formally, for two probability distributions, $p$ and $q$, where $p$ is a true distribution and $q$ is a given distribution over the same set of events, the cross-entropy measures the average number of bits required to identify an

event from a set of possibilities, if using $q$ rather than $p$. Basic intuition for using cross entropy is that unlikely events are regarded as more informative than likely events. For classification tasks in machine learning, $q$ takes the form of estimated class probabilities produced by a model, and $p$ are the corresponding correct class labels. For multi-class problems, the categorical cross-entropy loss function is given as,

$$\mathcal{L}(p, q) = -\sum_{x} p(x) \log q(x), \tag{4.8}$$

where, in this work, $q(x)$ is the output from the final softmax layer of the neural network, and $p(x)$ is the correct class labels. The outputs of the softmax function can be interpreted as posterior probabilities as, for classification problems where the desired outputs are zero or one, the cross-entropy cost function is minimised when the softmax outputs are posterior probabilities [Richard and Lippmann, 1991].

The correct and estimated class labels, $p(x)$ and $q(x)$, are encoded as "one-hot" vectors, where one-hot encoding allows for the representation of multiple classes in classification models. For example, given a task with four possible classes, an input example of the first class would have be assigned a one-hot vector of $[1, 0, 0, 0]$, an example of the second class would be assigned $[0, 1, 0, 0]$, and so on.

To prevent over-fitting of the model on the training data, the dropout technique [Srivastava et al., 2014], among other regularisation techniques, is used within the neural network architecture. During training, neurons are selected at random and dropped. That is, the neuron and its connections are temporarily removed from the network for that particular instance or set of training examples.

Figure 4.3a shows an example of a fully-connected neural network with a single hidden layer, whereas Figure 4.3b shows the same network after dropout has

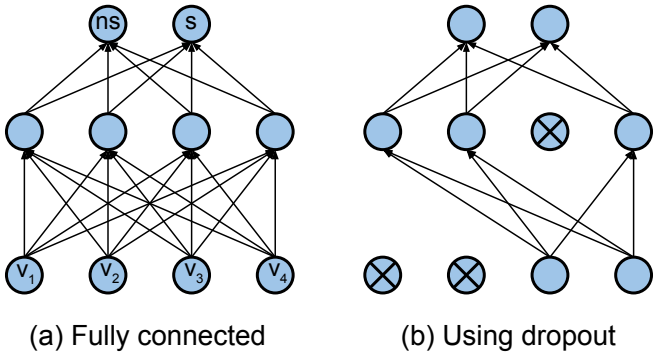



**Figure 4.3:** A fully-connected network is shown in (a), and the same network after dropout has been applied in (b).

been applied. A probability of $p = 0.5$ is typically used for dropout applied to fully-connected hidden layers, and a probability of zero, or close to, for dropping input units. The effect of applying dropout during training is to train a number of "thinned" models. For estimation, the classifications are then taken from the average of all the thinned networks. The effect is similar to training a large ensemble of models and averaging the predictions of each model [Goodfellow et al., 2013].

Training of the networks is performed using the resilient backpropagation algorithm [Riedmiller and Braun, 1993], with the primary benefit over the standard backpropagation algorithm being that training concludes considerably faster. The training visual vectors are grouped into mini-batches of 1024 examples, with $z$-score normalisation applied to the input 2D-DCT visual features. The weight values are initialised with uniformly distributed random variables in the range $-0.01$ to 0.01, and the learning rate is fixed at 0.001. The model training is completed once validation scores converge, and there is no further increase in the prediction accuracy of the models. More details of the neural network used are given in Appendix B.2.1.

Two neural network models are explored: `NN_DCT` that takes static 2D-DCT visual features as input, and `NN_DCT_`$\Delta$ that takes the static 2D-DCT visual features with the first and second order temporal derivatives as input.

### 4.3.2 Convolutional neural network

The convolutional neural network architecture is now explored for voicing classification, where the convolutional layers are used for visual feature extraction. The idea is that, instead of pre-extracting visual speech features using a 2D-DCT or an AAM, the model will extract suitable features from raw pixels located about the mouth of a speaker. Convolutional neural networks have shown application for myriad computer-vision tasks such as handwritten digit recognition and image classification, and are motivated by the function of the primary visual cortex [Murphy, 2012]. More recently, they have been successfully applied to the tasks of large-vocabulary continuous speech recognition (LVCSR) [Sainath et al., 2013], and have begun to be applied for audiovisual ASR [Noda et al., 2015].

Convolutional layers differ from fully-connected layers (as shown in Figure 4.3a) in that the units in layer $m$ are connected to only a local subset (representing a "receptive field") of the units in layer $m - 1$. Outputs from convolutional layers are named feature maps, and are calculated by convolving the inputs with multiple square matrices, which are analogous to filter kernels as used for image edge detectors or blurring. Weight sharing of the kernels ensures that features can be extracted independent of where they occur in the input.

An example CNN architecture is shown in Figure 4.4. In this example, an input image is convolved with four kernels producing four feature maps (shown in pink). A down-sampling stage (shown in green) is performed following the convolution stage to reduce the size (width and height) of the feature maps. Max-

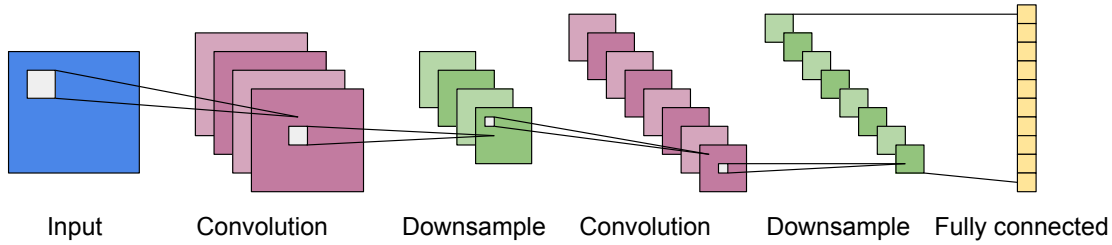



**Figure 4.4:** An example convolutional neural network architecture with two convolutional and down-sampling layers, connected to a final fully-connected output layer.

pooling is used to perform this sub-sampling, whereby the maximum output of a small square window is taken where the windows are non-overlapping. The size of the window determines the amount of down-sampling achieved, where, for example, a size of two will reduce the height and width of the feature maps by half. A further convolutional stage extracts eight feature maps, and is subsequently followed by another down-sampling layer. The output of the final sub-sampled layer is then input to a fully-connected layer. Stacking multiple convolutional and fully-connected layers together leads to the discovery of more higher-level global features [Sun et al., 2013].

The architecture used for this work follows Figure 4.4 and consists of two sets of convolution–max-pooling layers, followed by a fully-connected hidden layer, and a final output softmax layer. The first convolutional layer consists of thirty-two filters of size $3 \times 3$, and the second, sixty-four filters of size $3 \times 3$. Non-overlapping max-pooling follows each convolutional layer with square regions of size $2 \times 2$. The single fully-connected layer consists of 512 units, with dropout applied having probability $p = 0.5$. The window sizes and number of filters are chosen based on experiments conducted by Krizhevsky et al. [2012]. Rectified Linear Units are used throughout for the non-linear activation function. Further details of the CNN architecture used are given in Appendix B.2.2.

Training is performed using a graphics processing unit (GPU) card, which allows

for significantly quicker training over standard central processing units (CPU). The visual frame pixel intensities are scaled to be in the range of zero to one, and training is performed using mini-batches of size fifty. The network is trained using Nesterov's Accelerated Gradient Descent [Nesterov et al., 2007], with learning rate annealing performed, decreasing at a rate of 1 % per epoch. As with the standard neural network architecture, training is completed once validation scores converge and no further increase in classification accuracy is observed.

#### 4.3.2.1 Temporal information

When using deep neural networks for large-vocabulary continuous speech recognition applications, contiguous frames of audio vectors are concatenated to produce a single large feature vector to exploit longer-range temporal information [Hinton et al., 2012]. Applications where input data is a visual stream, such as objective video quality assessment and human action recognition, have led to the development of CNN architectures that incorporate temporal information. An approach using early-fusion has shown success in large-scale video classification [Karpathy et al., 2014], and is applied here.

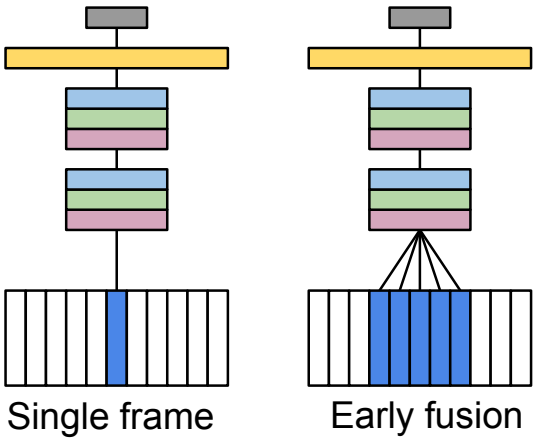



**Figure 4.5:** Static frame and early-fusion CNN architectures for including temporal information. Blue frames denote those that have current interest.

Figure 4.5 shows the single frame and early-fusion architectures, including the convolutional and fully-connected connections as shown in Figure 4.4. Early-fusion functions at the input layer, where the depth of the first convolutional layer filters is extended to convolve across neighbouring video frames. The idea being that the network will extract motion features including the visual feature representations. To explore the performance of this architecture, three video frames are stacked together, covering 30 ms of visual speech signal.

Experiments are conducted using two convolutional neural network systems, the first using a single frame input, `CNN_STATIC`, and the second using the early-fusion technique to stack three neighbouring frames together, called `CNN_STACK3`.

### 4.3.3 Aperiodicity estimation

The mapping from the class label, $\hat{c}_i^{\mathrm{VC}}$, to aperiodicity information is now described. To derive an aperiodicity surface using the voicing classification class labels, an aperiodicity output, $\mathbf{p}_i$, for each label from Equation 4.14 is required. To obtain an output for each label, the mean is taken from a set of training aperiodicity vectors, grouped by voicing class, resulting in $\boldsymbol{\mu}^v$ for voiced frames, $\boldsymbol{\mu}^u$ for unvoiced frames, and $\boldsymbol{\mu}^{ns}$ for non-speech frames. For example, given a voiced prediction from the voicing classification model, the aperiodicity vector, $\boldsymbol{\mu}^v$, will be used for that frame. The necessary mean aperiodicity vector can be output based on the estimated voicing class labels using

$$\mathbf{p}_i = \begin{cases} \boldsymbol{\mu}^v & \text{if } \hat{c}_i^{\mathrm{VC}} \text{ is voiced,} \\ \boldsymbol{\mu}^u & \text{if } \hat{c}_i^{\mathrm{VC}} \text{ is unvoiced ,} \\ \boldsymbol{\mu}^{ns} & \text{otherwise.} \end{cases} \tag{4.9}$$

Figure 4.6 shows the mean aperiodicity vectors for each class label for the female speaker extracted from the training data. As no speech exists during non-speech frames, the aperiodicity surface will be influenced by any noise and small misclassifications in speech boundaries, and is uniformly distributed in the range of $-5$ to $-10$ dB. The mean unvoiced vector has a lower intensity at the lower-frequencies, suggesting that even for unvoiced speech these components are still slightly periodic, and from 1 kHz onwards is around $-10$ dB, signifying more aperiodic components. The aperiodicity surface for the voiced frames exhibits a far lower intensity, indicating that the majority of components are strongly periodic. The increase in intensity to $-10$ dB at 4 kHz complies with the motivations behind harmonic plus noise models, where speech exhibits more noise-like characteristics at the higher-frequencies and can be incorporated into the models to yield a more natural speech output.

## 4.4 Aperiodicity estimation using joint feature clustering

The second method for aperiodicity estimation is presented using techniques from the area of vector quantisation, to produce codebooks of joint audio and aperiodicity features. The aperiodicity vectors are output given an input Mel-filterbank audio feature estimated from visual speech. Speech synthesis systems, such as HMM–TTS models, typically compress the aperiodicity surface into frequency bands [Silen et al., 2011]. Five frequency bands (0–0.5 kHz, 0.5 kHz–1 kHz, 1 kHz–2 kHz, 2 kHz–3 kHz, and 3 kHz–4 kHz) are used for audio with a sampling frequency of 8 kHz. To produce an output for each band, the mean of the frequency energies

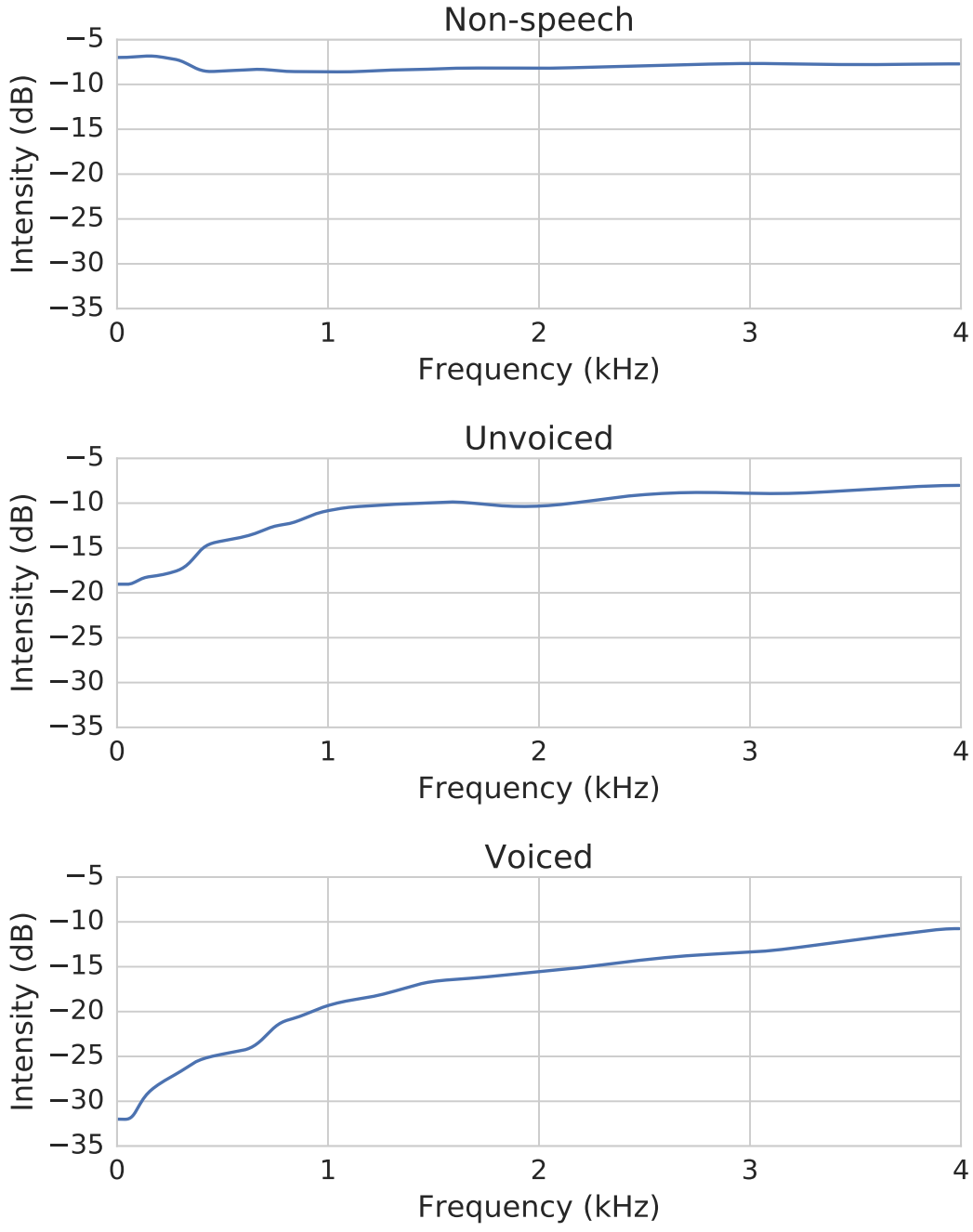



**Figure 4.6:** Mean aperiodicity vectors for the female speaker of the non-speech, unvoiced, and voiced class labels.

within each band is taken as follows

$$\mathbf{p}_i = \left[ p_i^{(0-0.5)}, p_i^{(0.5-1)}, p_i^{(1-2)}, p_i^{(2-3)}, p_i^{(3-4)} \right], \tag{4.10}$$

where $p_i^{(f_1-f_2)}$ is the aperiodic energy within the frequency band $f_1$ to $f_2$. Interpolation can be applied to reproduce a full aperiodicity surface in the range 0–4 kHz.

To produce the codebook, a joint clustering is performed on combined Mel-filterbank audio feature and band aperiodicity feature vectors, $\mathbf{z}_i$. The joint feature vectors,

$$\mathbf{z}_i = [\mathbf{a}_i, \mathbf{p}_i], \tag{4.11}$$

where $\mathbf{a}_i$ is a spectral-envelope audio feature vector, and $\mathbf{p}_i$ is a corresponding band aperiodicity feature vector, are extracted from the set of $N$ training features. The mini-batch $k$-means algorithm, discussed in greater detail in Chapter 6, is then applied to this set of joint training vectors, to produce a codebook, $C^{ap}$, with $K$ cluster centres, $\mathbf{c} \in C^{ap}$.

For aperiodicity estimation, given an estimated audio vector, $\hat{\mathbf{a}}_i$, (obtained using the visual-to-audio models in Chapters 6 and 7), the audio vector component of the codebook entries, $\mathbf{c}_j^A$, are searched and the aperiodicity component, $\mathbf{c}_{j^*}^P$, of the closest matching entry, $j^*$, is output using

$$\hat{\mathbf{p}}_i = \mathbf{c}_{j^*}^P \quad \text{where} \quad j^* = \arg\min_j \left\| \mathbf{c}_j^A - \hat{\mathbf{a}}_i \right\|^2, \tag{4.12}$$

and where $\hat{\mathbf{p}}_i$ is the estimated band aperiodicity feature vector. That is, as the joint feature vectors are modelled, by finding the cluster centre that has the lowest Euclidean distance between the spectral-envelope components, the aperiodicity

component of the selected cluster centre can be output.

## 4.5 Evaluation

To evaluate the performance of the methods proposed in this section for producing fundamental frequency and aperiodicity excitation parameters, each component is analysed separately. First, subjective intelligibility results of utterances reconstructed using the three artificial-$f_0$ methods are presented, with the spectral-envelope representations modelled using LPC coefficients and Mel-filterbank amplitudes. Secondly, classification accuracies for the voicing classification experiments are presented using the two neural network architectures, with an evaluation on the ability for convolutional neural networks to perform visual feature extraction. Finally, comparisons are made between the two aperiodicity estimation methods, voicing classification and joint-feature modelling, for producing aperiodicity surfaces.

### 4.5.1 Fundamental frequency

To evaluate the performance of the three artificial-$f_0$ methods for providing values of the fundamental frequency contour as required by STRAIGHT, subjective intelligibility experiments are conducted. Utterances from a male and female speaker from the GRID dataset are processed within an analysis-modification-synthesis framework, with the $f_0$ contour for each utterance modified using the artificial methods during reconstruction. The spectral-envelope representations are provided using two approaches for comparison: LPC coefficients and Mel-filterbank amplitudes. Utterances reconstructed using the original contour are also included, and form a baseline. The aim of these experiments is to determine what effect

the artificial contours have on intelligibility, whilst leaving the other parameters unchanged.

The intelligibility results described here are a subset of those from a larger experiment, where modifications are also explored for the spectral-envelope information. For reference, the subjective tests were conducted with twenty listeners, where each listener was presented with reconstructions of modified utterances from the GRID corpus (see Appendix A). Each utterance contains six words, with no context, and intelligibility is calculated as the word-level accuracy. More information on the experimental framework is provided in Section 5.2.

**Table 4.1:** Subjective intelligibility scores (and standard error) for the three artificial-$f_0$ methods plus the original, for the LPC coefficients and Mel-filterbank amplitudes spectral-envelope representations, for the female speaker.

| Audio feature | Original | Monotone | Time-varying | Unvoiced |
|---|---|---|---|---|
| LPC | 94.17 (1.78) | 94.17 (2.44) | 96.67 (1.90) | 96.67 (1.49) |
| Mel-filterbank | 94.17 (2.13) | 95.83 (2.00) | 94.17 (2.13) | 88.33 (3.35) |

The subjective intelligibility results, and standard error of the mean, calculated by,

$$\mathrm{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \tag{4.13}$$

where $\sigma$ is the standard deviation of the accuracy scores of all $n$ listeners, are reported in Table 4.1 for the female speaker. The scores are comparable for the majority of the different combinations of fundamental frequency contour type and spectral-envelope representation. Using the monotone $f_0$ contour resulted in the highest intelligibility (95.83%) for the Mel-filterbank amplitudes, and for the LPC audio features the time-varying and unvoiced contours were both best at 96.67%.

The corresponding male results are shown in Table 4.2, where, again, the majority of combinations are all similar. The intelligibilities are generally higher

**Table 4.2:** Subjective intelligibility scores (and standard error) for the three artificial-$f_0$ methods plus the original, for the LPC coefficients and Mel-filterbank amplitudes spectral-envelope representations, for the male speaker.

| Audio feature | Original | Monotone | Time-varying | Unvoiced |
|---|---|---|---|---|
| LPC | 96.67 (1.90) | 95.83 (1.61) | 95.83 (1.61) | 98.33 (1.12) |
| Mel-filterbank | 96.67 (1.90) | 92.50 (2.49) | 97.50 (1.33) | 97.50 (1.33) |

overall than for the female speakers. The time-varying and unvoiced artificial-$f_0$ contours gave the best scores (97.50%) for the Mel-filterbank features, and the unvoiced method was the best contour for the LPC audio features with an accuracy of 98.33%. For both audio feature representations, the unvoiced artificial contour yields the highest accuracies for both speakers, except for the female speaker when using Mel-filterbank amplitudes.

Importantly, the subjective intelligibility scores recorded show that the artificial fundamental frequency methods are not having an adverse effect on the intelligibility of the reconstructed utterances, which is important for the thesis of reconstructing intelligible audio speech using visual speech information.

### 4.5.2 Voicing classification accuracy

To determine the ability of the voicing classification system to produce a suitable aperiodicity surface output, it is necessary to first determine the classification accuracies for the system. Accordingly, accuracy results are presented for voicing classification, and voice activity detection, of the neural network and CNN systems, including a GMM method as a baseline. Accuracies are recorded for the multi-class voicing classification task, and then by grouping the unvoiced and voiced estimations, the voice activity detection results are obtained.

To measure voicing classification, reference labels are required for each frame of

speech, $i$, classifying each as either non-speech, unvoiced, or voiced. Voice activity labels are first used to set frames as either non-speech or speech. The PEFAC pitch-extraction algorithm [Gonzalez and Brookes, 2014] is then used to provide a probability that a given frame of speech is voiced. A threshold is applied to the voiced speech probabilities output from PEFAC, with speech fames having probability $p(t) \geq 0.5$ labelled as voiced, and frames with $p(t) < 0.5$ labelled as unvoiced. Voicing class labels can then be assigned to frames using

$$c_i^{\mathrm{VC}} = \begin{cases} \mathrm{v} & \text{if speech and } p(i) \geq 0.5, \\ \mathrm{u} & \text{if speech and } p(i) < 0.5, \\ \mathrm{ns} & \text{otherwise.} \end{cases} \tag{4.14}$$

The baseline GMM and standard neural network system uses 2D-DCT visual features as input, whereas for the CNN system raw pixel-intensities are provided. The video data is up-sampled to 100 Hz to match a typical audio speech frame rate, with each video frame converted to greyscale. Matrices of size $96 \times 96$ pixels are extracted about a centre-point of the speakers mouth, calculated from landmark data, and resized to $64 \times 64$ pixels. For the `CNN_STACK3` configuration, the greyscale matrices from three contiguous frames are stacked centred on a middle frame, producing a three-dimensional matrix with dimensions of $64 \times 64 \times 3$.

#### 4.5.2.1 Baseline model

In Almajai and Milner [2008], Gaussian mixture models are used to model visual feature vectors for the task of visual-only voice activity detection. In this work, a similar idea is applied to form the baseline against which to compare the neural network and CNN systems. A more detailed review of Gaussian mixture models

is given in Section 5.3.1. Vectors are grouped by class label and individual GMMs are trained: $\Phi^{\text{ns}}$ for non-speech frames, $\Phi^{\text{u}}$ for unvoiced frames, and $\Phi^{\text{v}}$ for voiced frames. Classification is performed by taking the arg max of the probabilities produced by each class GMM, $\Phi^l$, given the input visual vector, $\mathbf{v}_t$, using

$$\hat{c}_t^{\text{VC}} = \arg\max_l \left( p(\mathbf{v}_t|\Phi^l) \right), \tag{4.15}$$

where $l \in \{\text{ns, u, v}\}$. Through experimentation it was found that using sixteen clusters for each GMM gave the best performance.

The two GMM models are named `GMM_DCT` and `GMM_DCT_`$\Delta$, for the static and temporal models respectively. The voice activity detection results are obtained by grouping the unvoiced and voiced class labels to give a speech/non-speech decision.

#### 4.5.2.2 Experiment results

The experiments described here are used to evaluate the classification accuracy of the three models: GMM, NN, and CNN. Each of the models are trained on input visual data, either 2D-DCT or greyscale pixel intensities, to predict the voicing class labels. From the female speaker, 800 utterances are used for training, and 200 utterances are used for testing. The accuracy is determined by recording the number of correct class predictions on the test data.

Table 4.3 shows voicing classification and voice activity detection accuracies for the female speaker. The `CNN_STACK3` achieves the best accuracy for voicing classification, with 87.55 % of frames classified correctly. Accordingly, the same system outperforms both the GMM and neural network systems for voice activity detection with an accuracy of 97.66 %. Surprisingly, the `CNN_STATIC` system is able to achieve 86.05 % voicing accuracy using static information. In comparison, the

static GMM and neural network systems achieve accuracies of 11.76 % and 6.44 % lower, respectively. This suggests that by using convolutional neural networks, suitably descriptive visual speech feature representations can be found. Furthermore, and as is to be expected, the accuracies achieved for voice activity detection are higher than the voicing classification scores.

**Table 4.3:** Voicing classification and voice activity detection accuracies in per cent.

| Configuration | Voicing accuracy | VAD accuracy |
|---|---|---|
| GMM_DCT | 74.29 | 92.61 |
| GMM_DCT_$\Delta$ | 78.99 | 94.34 |
| NN_DCT | 79.61 | 96.00 |
| NN_DCT_$\Delta$ | 86.35 | 96.80 |
| CNN_STATIC | 86.05 | 96.99 |
| CNN_STACK3 | **87.55** | **97.66** |

Increased voicing classification accuracy by including temporal information is readily apparent for both the neural network and GMM systems. A classification accuracy increase of 4.7 % and 6.7 % is gained for the GMM and neural network respectively. However, the same increase does not occur when using the CNN. Interestingly, it appears that due to the only slight increase in performance between the CNN_STATIC and CNN_STACK3 systems of 1.5 %, using the early-fusion technique for including temporal information in the CNN architecture is not ideal for this work. Accordingly, other techniques of incorporating temporal information, such as recurrent neural network architectures with convolutional layers [Donahue et al., 2015], could result in a greater accuracy as they are better able to exploit longer-range dependencies in the data.

Table 4.4 shows a confusion matrix for voicing classes predicted by the CNN_STACK3 model. The majority of voicing classification errors occur with the misclassification of unvoiced frames as voiced frames, with 27.25 % doing so. The problem experienced with voicing classification occurs when different voiced and unvoiced

**Table 4.4:** Confusion matrix of per cent classification accuracy using the `CNN_STACK3` model.

| | Non-speech | Unvoiced | Voiced |
|---|---|---|---|
| Non-speech | 98.23 | 1.49 | 0.28 |
| Unvoiced | 5.91 | 66.84 | 27.25 |
| Voiced | 0.72 | 8.93 | 90.36 |

phonemes have the same visual speech realisations. Phonemes sharing the same visual realisations can be grouped by phoneme equivalence class (PEC), a generalisation of the viseme. The idea behind PECs, proposed by Auer Jr and Bernstein [1997], is that similar visual realisations of different phonemes can be grouped together into the same class. Regarding the problem of voicing classification, a PEC comprised of /s t z/ consists of two unvoiced consonants, /s/ and /t/, and a voiced consonant, /z/, for example. A PEC comprised of /f v/ has a voiced and unvoiced consonant. Voice activity detection errors can be seen where unvoiced or voiced frames are classified as non-speech, and vice versa. The problem in this case is that visual realisations of certain PECs have a mouth shape that is very visually similar to the neutral. For example, this is the case with the PEC comprised of the phonemes /b m p/, where the mouth tends to be closed. The majority of errors occur when unvoiced frames are misclassified as non-speech frames, happening for 5.91 % of unvoiced frames.

This set of experiments on comparing a baseline Gaussian mixture and neural network models using DCT visual features, and a CNN approach using non-engineered raw visual frames, was intended to show that the CNN systems gave results comparable to more typical approaches that used pre-extracted, engineered visual features. The CNN system contains a number of convolution and max-pooling layers functioning as feature-extractors, the output of which are fed into a fully-connected hidden layer before the output softmax layer. When this area

of work was published in Le Cornu and Milner [2015], it was in an attempt to motivate further exploration into the technique. Using a CNN for visual-feature extraction is now a common technique and numerous papers have been published using this approach with great success [Noda et al., 2015; Assael et al., 2016; Chung et al., 2016].

### 4.5.3 Codebook size for joint aperiodicity estimation

For producing accurate aperiodicity estimations using the joint-feature codebook approach, it is first necessary to evaluate various codebook sizes. To find an optimal number of cluster centres, $K$, to use in the aperiodicity codebook, $C^{ap}$, the mean squared error is recorded for audio-only comparisons between the original band-aperiodicity feature vectors and their quantised counterparts with various codebook sizes. The quantised aperiodicity features are output using Equation 4.11 given the original Mel-filterbank audio features as input.

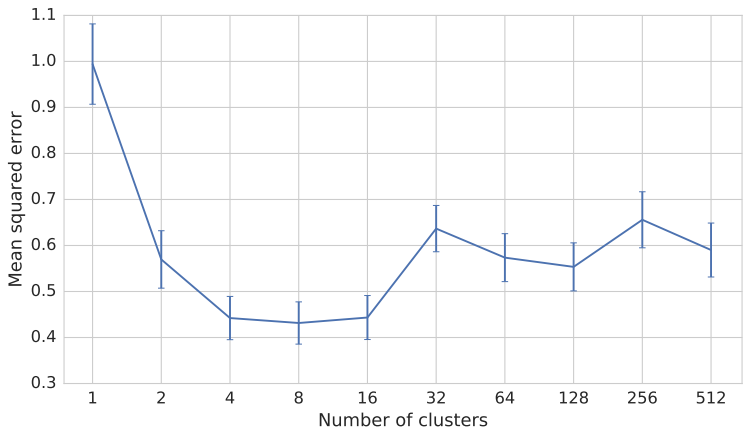



**Figure 4.7:** Mean squared error (with error bars showing a single standard error) between original and quantised band-aperiodicity features with increasing codebook size, $K$.

In Figure 4.7, the MSE is shown with respect to codebooks with increasing num-

bers of clusters, $K$. The MSE decreases quickly as the codebook size is increased to $K = 4$, with the optimal number of codebook entries found at $K = 8$, and then increases again from $K = 16$ onwards. This effect confirms informal listening tests that suggest codebooks with fewer entries are beneficial, as aperiodicity surfaces derived from codebooks with a large number of entries exhibit extremely erratic frame-by-frame changes, resulting in adversely affected reconstructed audio speech.

### 4.5.4 Aperiodicity estimation

Experiments conducted on aperiodicity estimation using the two methods presented in this chapter are now evaluated. The first method utilises a voicing classification system to predict voicing class labels, where mean aperiodicity vectors can then be output based on whether the label is non-speech, unvoiced, or voiced. The second method uses techniques from the area of vector quantisation to produce a joint Mel-filterbank and band aperiodicity codebook, from which an aperiodicity vector can be obtained by searching for the codebook entry to which the input Mel-filterbank vector is closest.

Experiments are conducted for both audio-only and visual-to-audio configurations. For the voicing classification method, the audio-only experiments are conducted by producing the three mean aperiodicity vectors (non-speech, unvoiced, and voiced) from the training data, and then using the ground-truth voicing labels from the test set to produce the quantised aperiodicity surfaces. For the joint-feature method, the codebook is built using the training data, and then the ground-truth Mel-filterbank vectors from the test set are used to select the appropriate aperiodicity vectors. Conversely, the visual-to-audio experiments explore the aperiodicity estimation as if the models where being used as part of a fi-

nal system. For the voicing classification method, the `CNN_STACK3` model is used to predict the voicing labels on the test set from input visual features, with the mean aperiodicity vectors output based on the label. Whereas for the joint-feature method, estimates of the Mel-filterbank audio features are obtained using the best performing visual-to-audio model from Chapter 6.

**Table 4.5:** Mean squared error for the two proposed aperiodicity estimation methods for both audio-only and visual-to-audio scenarios.

| Method | Audio-only | Visual-to-audio |
|---|---|---|
| Voicing classification | 0.717 (0.079) | 0.770 (0.097) |
| Joint aperiodicity | 0.561 (0.058) | 0.626 (0.096) |

Table 4.5 shows the MSE between the original and estimated aperiodicity features for the audio-only and visual-to-audio configurations for each of the two systems. It is readily apparent that the joint-feature method achieves lower MSEs over the voicing classification method, for both configurations. For the audio-only scenario, the joint-feature method shows an MSE of 0.561, in comparison to 0.717 when using voicing classification, indicating that even with ground-truth test data the aperiodicity estimates are poorer using the later system. This observation is further confirmed in the visual-to-audio scenario, where the joint-feature method shows an MSE of 0.626, in comparison to 0.770 when using voicing classification. As the voicing classification method relies on accurate class label predictions for aperiodicity estimation, any errors in predicting the voicing labels will result in a higher MSE. Accordingly, given the lower MSEs recorded, the joint-feature method is used for producing aperiodicity estimations for the remainder of the work conducted in this thesis.

## 4.6 Summary

In this chapter, various methods are proposed for generating the necessary excitation information as required by the STRAIGHT speech production model. Three artificial methods are developed for producing fundamental frequency contours, including two methods of aperiodicity estimation.

The subjective intelligibility experiments conducted on the three artificial-$f_0$ methods show that, for accurate spectral-envelope representations, the contours do not adversely affect the intelligibility of the reconstructed audio. For both the LPC coefficients and Mel-filterbank amplitude spectral-envelope representations, the audio is highly intelligible for both the male and the female speakers, across all three artificial methods. Accordingly, their use is explored further in the following chapter to see what effect they have when used for spectrally smoothed speech.

The work on voicing classification, and voice activity detection, has shown that frame-level accuracies of 88 % and 98 % result for voicing classification and VAD, respectively. The convolutional neural network approach outperforms the baseline GMM and neural network systems for both voicing classification and voice activity detection, where the high accuracy achieved for the CNN using static information shows promise for their ability to discover descriptive visual speech feature representations.

The experiments conducted on aperiodicity estimation have shown that the joint-feature method outperforms the voicing classification method for both audio-only and visual-to-audio configurations, and is used throughout the remainder of this work for audio speech reconstructions using STRAIGHT. Although the voicing classification method performs poorly for aperiodicity estimation, in comparison to the joint-feature method, the technique still has application for voice activity detection using visual speech.

In the following three chapters, work is presented on producing spectral-envelope estimates from visual speech information, which can be used for reconstructing intelligible audio speech signals using STRAIGHT.

# Chapter 5

# Regression system

## 5.1 Introduction

Two methods are explored in this thesis for producing spectral-envelope estimates from visual speech input, the first approach, detailed in this chapter, explores using regression methods for this task. That is, given the input visual features, the mapping models are used to estimate real-valued and continuous audio feature coefficients. The audio feature representations of the spectral-envelope and visual feature representations of the visual articulators have been discussed previously in Chapter 3. An overview of the system configuration explored in this chapter is shown in Figure 5.1. Statistical models commonly used in many areas of speech processing are Gaussian mixture models and deep neural networks. Both have had successful application for tasks such as large-vocabulary continuous speech recognition [Sainath et al., 2013], speech enhancement [Xu et al., 2014], and text-to-speech (TTS) synthesis [Qian et al., 2014]. Additionally, both models can be configured to perform regression [Xu et al., 2015; Park and Kim, 2000].

In this Chapter, various system configurations are explored with different com-

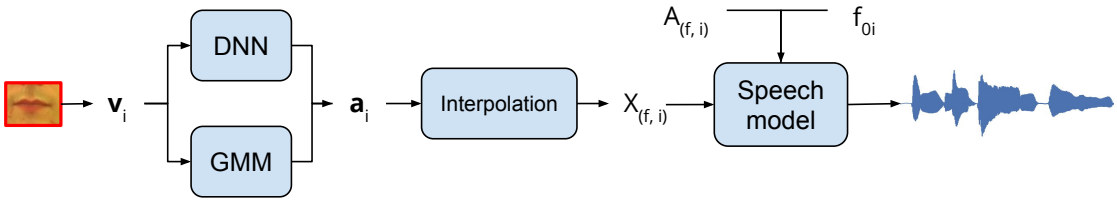



**Figure 5.1:** Regression system overview. Visual features are extracted from the mouth of a speaker and input to the visual-to-audio regression mapping models, outputting audio feature estimates. Interpolation is applied to produce spectral-envelopes, which are input to a speech production model along with artificial excitation to reconstruct audio speech.

binations of audio and visual feature representations, and with the two types of model. Furthermore, subjective experiments are presented on audio speech reconstructed from reduced-dimensionality audio features. This is motivated by the idea that it may be beneficial to estimate feature vectors with fewer coefficients, resulting in greater overall intelligibility. The effect of reducing the dimensionality of the audio features introduces a smoothing effect on the spectral-envelope. This smoothing effect is explored initially in a controlled manner to determine whether reduced dimensionality features can still yield sufficiently intelligible speech reconstructions. Their use within the proposed visual-to-audio framework is then explored further, to see if the hypothesis is valid.

The remainder of this Chapter is organised as follows. In Section 5.2, the spectral smoothing experiments are discussed. For obtaining audio domain estimates from visual domain information, two types of models are explored in Section 5.3, specifically: multivariate Gaussian mixture models and deep neural networks. Results from objective and subjective intelligibility experiments, including an analysis of reconstructed audio utterances, are presented in Section 5.4. Lastly, an overview of this section of work is given in Section 5.5.

## 5.2 Spectral smoothing

Smoothing of the spectral-envelope, by using audio feature representations with fewer coefficients, is explored with the expectation that it may be beneficial for the visual-to-audio mapping model to estimate fewer coefficients with greater accuracy. The hypothesis being that audio features with fewer coefficients, i.e. a lower dimensionality, can be estimated with a greater overall accuracy than audio features with a larger number of coefficients, and that the resultant reconstructed audio speech has greater intelligibility over the features with a higher dimension, yet potentially poorer accuracy. For example, LPC audio features with order $P = 4$ estimated with a total accuracy of, say, 90 % may result in speech with an intelligibility of 80 %, as opposed to LPC features with order $P = 10$ estimated with an accuracy of 75 %, resulting in reconstructed audio speech with an intelligibility of less than 80 %.

The effect of smoothing the spectral-envelope is similar to that of smearing the spectral information in the frequency domain. That is, to reduce the intensity of the spectral peaks and to raise the intensity of the spectral troughs, resulting in a flatter spectrum with increased formant bandwidths. The overall energy of the spectral-envelope remains the same, where the energy of the frequencies with greater amplitudes is distributed to the neighbouring frequencies with less amplitude. This effect is observed for certain types of hearing loss where impairments are attributed to broadening of the auditory filters in the ear, resulting in lower speech intelligibility [Baer et al., 1993].

To achieve this smoothing effect, the number of audio feature coefficients used to represent the spectral-envelope information for each frame is reduced. From Chapter 3, the audio features explored in this work are linear predictive coding (LPC) coefficients and Mel-filterbank channel amplitudes. The amount of smooth-

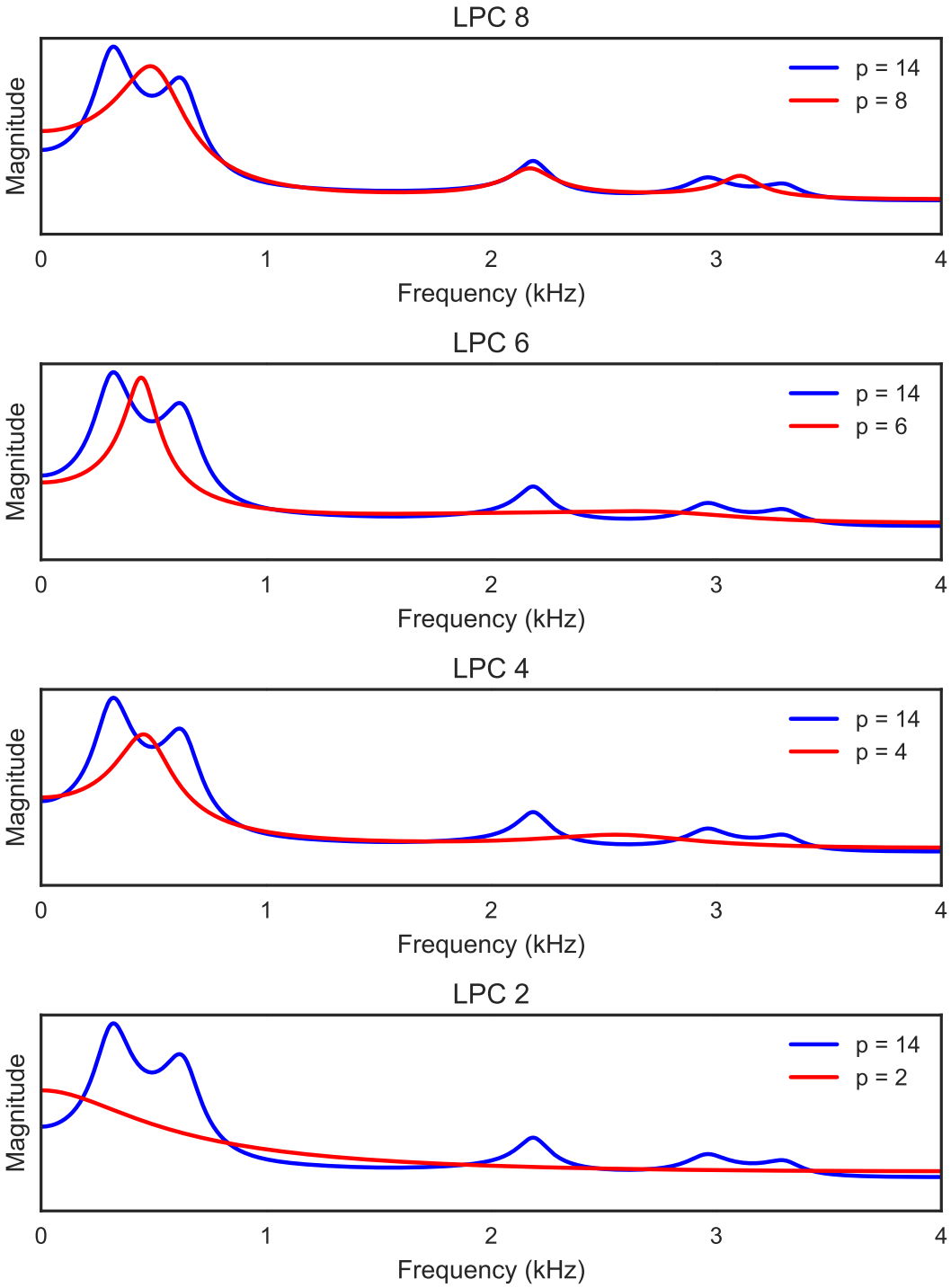



**Figure 5.2:** Comparison of LPC features with increasing levels of smoothing applied. The red lines show the smoothed spectral-envelopes, whereas the blue lines show the spectral-envelope of a frame using LPC features with an order of $P = 14$.

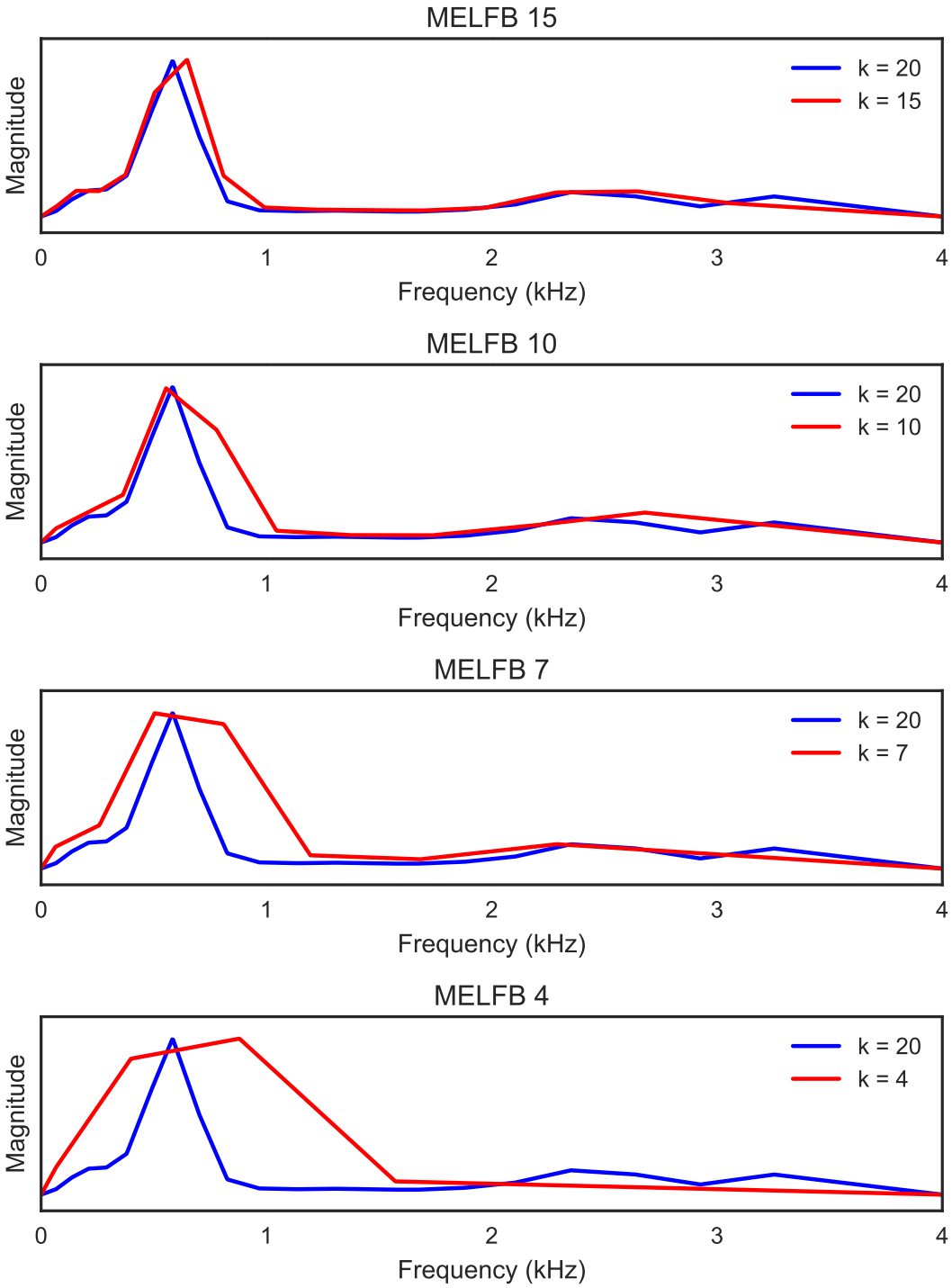



**Figure 5.3:** Comparison of Mel-filterbank features with increasing levels of smoothing applied. The red lines show the smoothed spectral-envelopes, whereas the blue lines show the spectral-envelope of a frame using Mel-filterbank features with a channel number of $K = 20$.

ing applied is varied by using orders of $P = \{2, 4, 6, 8, 14\}$ for the LPC features, and filterbank sizes of $K = \{4, 7, 10, 15, 20\}$ for the Mel-filterbank features. The effect of applying increasing amounts of smoothing for a single frame is shown in Figure 5.2 for LPC, and for Mel-filterbank in Figure 5.3. An order of $P = 2$, and a filterbank size of $K = 4$, give the greatest amount of smoothing, with a noticeable reduction in intelligibility from informal listening tests. The result of smoothing is that there is a loss of formant structure in the reconstructed speech. Conversely, speech reconstructed from the audio features with the greatest order and filterbank size, for LPC and Mel-filterbank respectively, is near indistinguishable from the original utterances in terms of both intelligibility and quality. Spectrograms for the utterance "lay blue at j 6 please" spoken by a female speaker, and reconstructed from LPC features with smoothing applied are shown in Figure 5.4, and from Mel-filterbank features in Figure 5.5.

Ultimately, for the hypothesis of producing intelligible speech from audio features with reduced parameters to be valid, it is necessary to determine the intelligibility of reconstructed audio speech utterances with the various levels of smoothing applied. To explore this, audio speech utterances are processed within an analysis-modification-synthesis (AMS) framework. Using the AMS framework, speech parameters are extracted during an analysis stage with modifications applied to the parameters through smoothing of the spectral-envelope and use of the artificial-$f_0$ methods. The modified parameters are provided as inputs to reconstruct audio utterances in the speech synthesis stage. Subjective listening tests are then conducted to determine intelligibility scores. The GRID audiovisual corpus is used for this set of experiments, and is detailed in Appendix A.1.

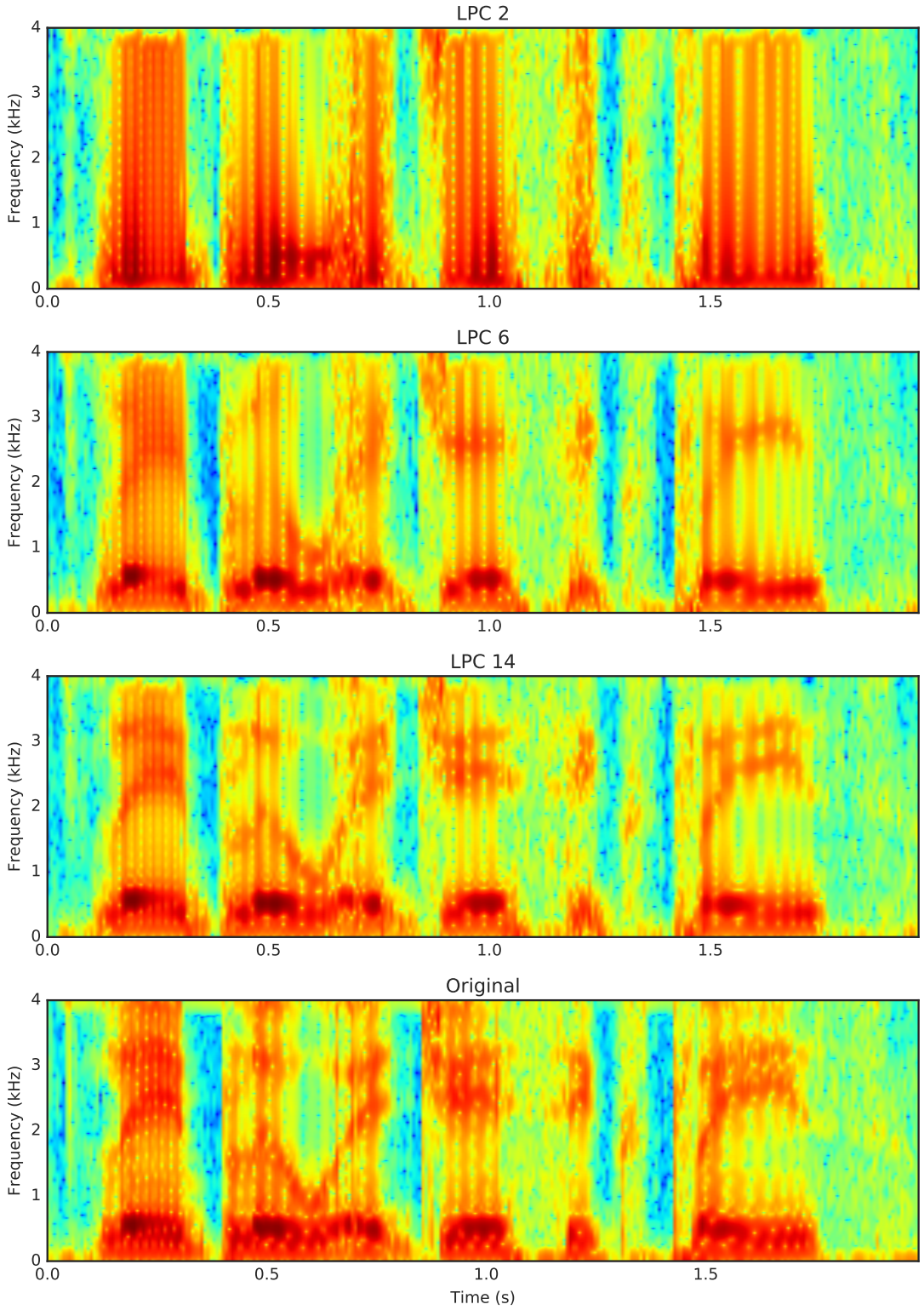



**Figure 5.4:** Wideband spectrograms of utterances reproduced from LPC features with spectral smoothing applied. The original utterance is included for comparison.

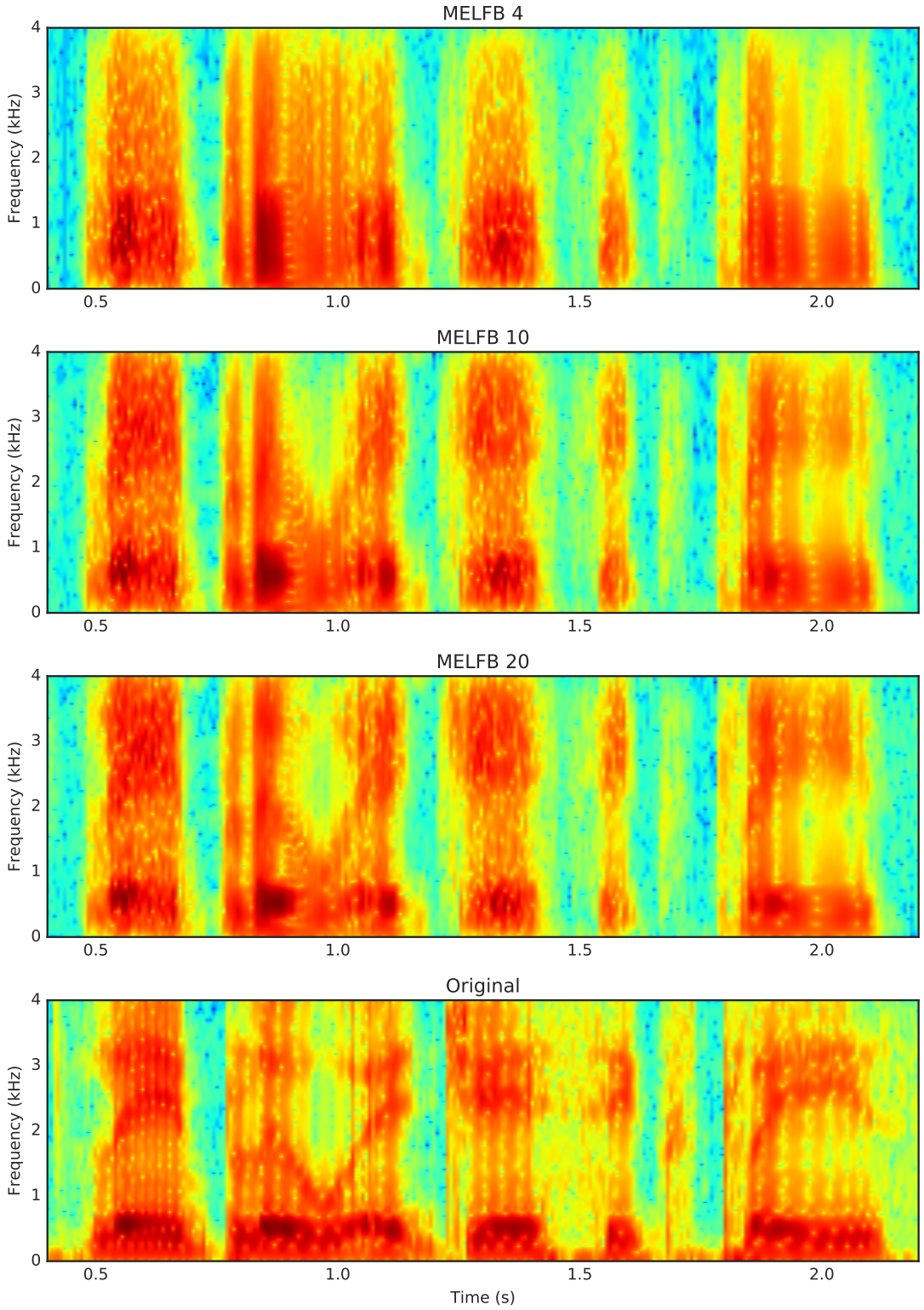



**Figure 5.5:** Wideband spectrograms of utterances reproduced from Mel-filterbank features with spectral smoothing applied. The original utterance is included for comparison.

### 5.2.1 Subjective tests

Subjective listening tests were conducted with twenty listeners to determine how the intelligibility of audio utterances is affected through increased application of spectral-envelope smoothing. Each subject is presented with 82 different utterances, of which 41 are from a female speaker and 41 from a male speaker. The 41 utterances comprise the three artificial-$f_0$ methods, as discussed in Chapter 4, plus reproductions with the original (ground-truth) fundamental frequency contours. For the audio representations, LPC features with orders of $P = \{2, 4, 6, 8, 14\}$, and Mel-filterbank channel amplitudes with channel numbers of $K = \{4, 7, 10, 15, 20\}$ were used. An unprocessed utterance from the corpus was included for each speaker to obtain a baseline level of intelligibility.

| command | colour | preposition | letter | digit | adverb |
|---|---|---|---|---|---|
| bin | blue | at | A-Z | 1-9 | again |
| lay | green | by | minus W | zero | now |
| place | red | in | | | please |
| set | white | with | | | soon |

**Table 5.1:** GRID sentence grammar, with available choices per word.

For each utterance, the subjects were asked to select which of the available word choices, as shown in Table 5.1, for each component of the grammar:

<command> <colour> <preposition> <letter> <digit> <adverb>,

they believed to be correct. Therefore, accuracy was calculated on a per-word basis. To remove any bias that could affect the results, utterances were presented to the subjects in a random order, and for each of the 82 utterance types there was a choice of twenty utterances, with one utterance being selected at random. That is, assuming twenty listeners took part, on average each listener would hear

an utterance unheard by the other listeners. Furthermore, utterances were not replaced, and so a listener would only here a particular utterance once.

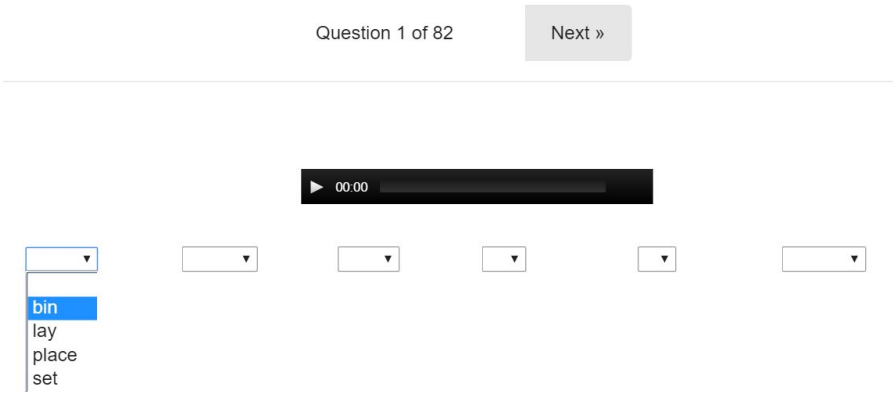



**Figure 5.6:** Screen capture of a question page of the web-based subjective experiment test interface. In this example, an audio file can be listened to with the word choices presented in the selection boxes below.

Testing was conducted using a web-based interface, as can be seen in Figure 5.6, with the utterances displayed in an audio player. The subjects could listen to each utterance as many times as they desired. To select their word choices, six drop-down selection boxes were displayed. Subjects who did not complete the test in a sound-proof room were asked to situate themselves in a quiet environment using a pair of high-quality headphones, and to complete the test in one sitting.

### 5.2.2 Evaluation

Results were averaged across all listeners for each utterance configuration and were scored based on word accuracy, giving a per cent intelligibility score for each of the 82 configurations. Figure 5.7 shows the intelligibility scores for combinations of each of the artificial fundamental frequency methods with the LPC and Mel-filterbank audio features, where the male and female speaker results have been grouped.

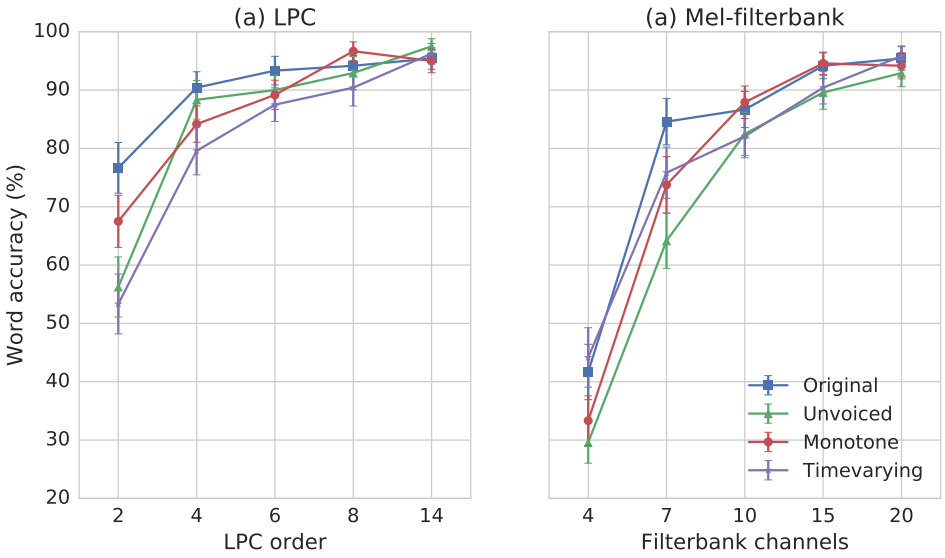



**Figure 5.7:** Intelligibility scores (with error bars showing a single standard error) for the various combinations of artificial-$f_0$ method and audio features, with various levels of smoothing applied. Results from the male and female speakers have been grouped by audio feature type.

The results in Figure 5.7 show that using LPC coefficients results in better intelligibility than using Mel-filterbank amplitudes for similar numbers of features. It is believed the greater accuracy for LPCs is due to the spectral-envelope fitting the spectral peaks better than the Mel-filterbank channels. If the Mel-filterbank channels are not located at the spectral peaks then a further distortion of the original spectral-envelope is being introduced, adversely affecting the reproduced utterances. Interestingly, using only two LPC coefficients gives an accuracy of 77 % when using the original $f_0$, and 68 % when using the monotone contour. The level of spectral detail retained using only two LPC coefficients (see Figure 5.4) is extremely low, yet there is evidently enough information to achieve relatively high accuracy scores. Increasing the number of coefficients to four results in an average accuracy of roughly 85 % across all artificial-$f_0$ methods.

A comparison of the artificial-$f_0$ for the LPC audio features shows that the time-

varying contour achieves the lowest accuracy for all numbers of coefficients except fourteen. With two coefficients, the monotone contour is the best artificial method, with an absolute accuracy over 10 % greater than the next best artificial method: unvoiced excitation. However, with four coefficients the unvoiced method is around 4 % higher than the monotone, and comparable to the accuracy achieved using the original contour. With six coefficients the accuracies achieved are roughly equal, yet with eight coefficients the monotone method manages to outperform even the utterances with the original $f_0$ contour. With fourteen coefficients, all artificial-$f_0$ methods differ by only a few per cent.

Comparing the artificial-$f_0$ methods for the Mel-filterbank audio features shows that the unvoiced excitation performs the worst across all numbers of channels. With four channels the time-varying and original methods achieve similar accuracies of around 43 %. With seven channels the performance of the monotone method increases and becomes comparable with the time-varying method, although the original method is best overall. With channel numbers of ten and fifteen, the monotone and original methods are equivalent, at roughly 86 % and 94 % respectively. With twenty channels the artificial method results differ by roughly 4 %, with the time-varying contour achieving the best accuracy. The unvoiced, time-varying, and original results, are still increasing from fifteen to twenty channels, suggesting that a greater number of channels may result in a further increase in intelligibility.

In Figure 5.8, the intelligibility scores are divided between the two audio feature types for the individual male and female speakers. The intelligibility scores are comparable across all LPC orders for both the female and male utterances. There appears to be no general trend for one artificial excitation method to outperform the others, as the performance of the methods differs for each LPC order. For both genders, the time-varying contour typically under-performs the other artificial-$f_0$

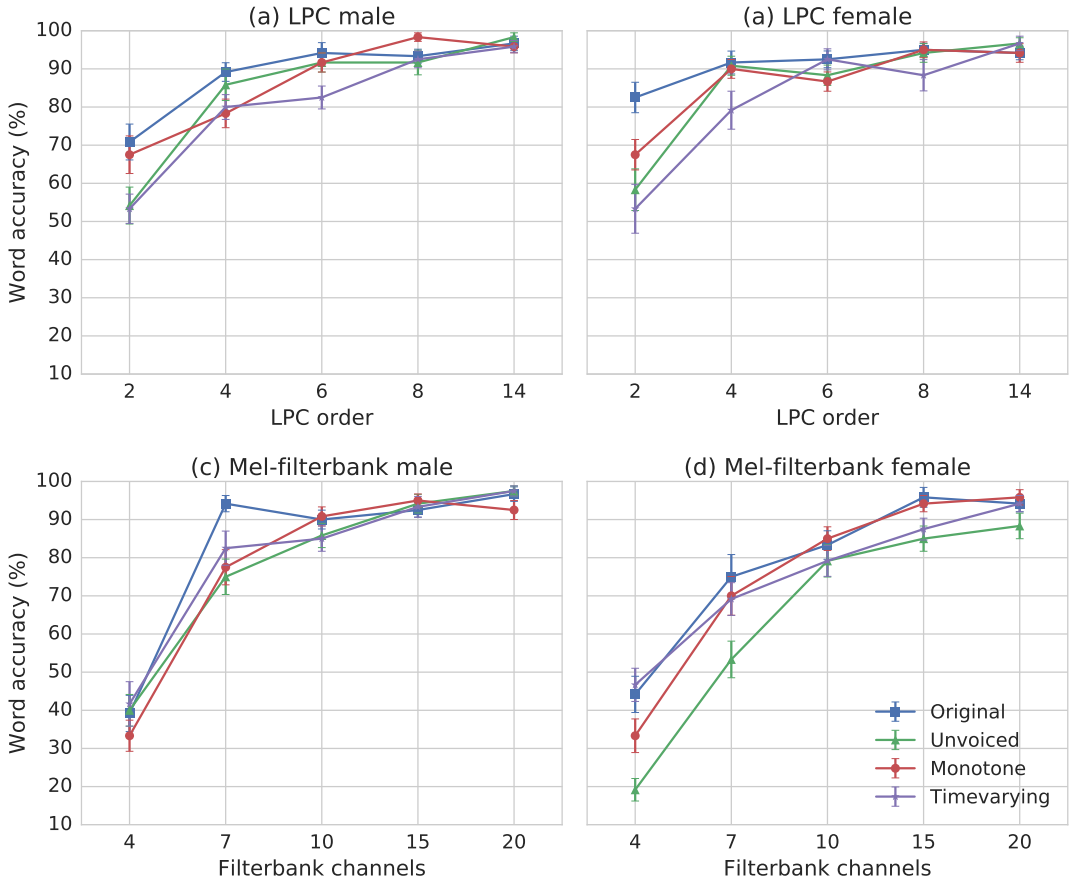



**Figure 5.8:** Comparison of male and female intelligibility scores for LPC and Mel-filterbank audio features.

methods, suggesting that the contour itself is affecting the intelligibility of the reproduced speech more so than the other methods. Surprisingly, using the original excitation with only two LPC coefficients for the female speaker, an accuracy of 82.5 % results.

For the Mel-filterbank audio features, the intelligibility scores are similar with four channels for both genders, except for the female speaker using unvoiced excitation. With seven and ten channels the male scores are greater than the scores for the female, and at fifteen and twenty they are similar. The tendency for the unvoiced excitation to be worse for the female occurs across all numbers of filterbank channels. The male result for Mel-filterbank with seven channels using the original excitation appears to be an anomaly with a score of 94.17 %. This is significantly higher than the other artificial methods for this number of channels, and greater than the scores achieved for all $f_0$ methods for both ten and fifteen channels, except monotone excitation with fifteen channels.

A known problem with the LPC model is that it is unable to model the nasal cavity [Kang and Lee, 1988]. An informal experiment comparing confusion matrices of the letters between the LPC and Mel-filterbank utterances shows no obvious evidence of this drawback. Confusions exist between the nasal consonant phonemes /m/ and /n/ for both audio feature representations. Given the already poorer intelligibility of the reconstructed utterances when compared to the unaltered equivalents, the intelligibility will likely only be marginally affected by this deficiency.

In summary, of the two audio features, the LPC features give speech reproductions with greater intelligibility over Mel-filterbank amplitudes when using fewer features. With regards to the artificial-$f_0$ contours, for LPC features the monotone and unvoiced methods perform best, and for Mel-filterbank amplitudes, the mono-

tone method outperforms all of the others. The female speech reproductions result in greater intelligibility with a reduced number of features. This suggests that initial experimental work for the visual-to-audio project should first focus on a highly intelligible female speaker, and then on a male speaker. Additionally, using the original fundamental frequency contour results in greater intelligibility scores over the artificial methods, for the majority of audio feature configurations, indicating the importance of having the correct $f_0$ contour, and accordingly, voicing.

## 5.3 Visual-to-audio mapping models

Having determined how spectral smoothing affects the intelligibility of speech, and the effect introduced by the artificial-$f_0$ methods, the next step is to explore the models for performing the visual-to-audio domain mapping for estimating spectral-envelope from visual speech. Various machine learning approaches and statistical techniques can be explored for producing spectral-envelope estimates from input visual features. Two commonly used probabilistic models with wide-ranging application in speech processing are multivariate Gaussian mixture models and deep neural networks. Only a brief overview of the specific details of each of the two models is given in this section, for a more detailed treatise the reader is encouraged to consult the work of Murphy [2012].

### 5.3.1 Gaussian mixture models

Gaussian mixture models (GMMs) are a type of probabilistic model that have been used for a number of decades in various applications of audio and audiovisual speech processing. For speech recognition tasks, GMMs were the *de facto* method, until the recent advances and successful employment of deep neural networks for

performing acoustic modelling. As with all mixture models, the underlying data is assumed to belong to a mixture distribution, and in the case of GMMs, the individual distributions can be modelled using a multivariate Gaussian.

To perform the mapping from the visual to the audio domain, a GMM is created to model the joint density of audio and visual features from an individual speaker. A joint audiovisual feature vector, $\mathbf{z}_i$ is produced by augmenting audio feature vectors with their corresponding visual feature vectors,

$$\mathbf{z}_i = [\mathbf{a}_i, \mathbf{v}_i]\,, \tag{5.1}$$

where $\mathbf{a}_i$ and $\mathbf{v}_i$ correspond to the $i$th audio and visual feature vectors, respectively. The dimensionality of the joint feature vector is the summation of the dimensions of the individual feature vectors. To build a joint audiovisual GMM, $\Phi^{av}$, the expectation-maximisation (EM) algorithm is applied to a training set of joint audiovisual features. To initialise the algorithm, the standard Lloyd's $k$-means algorithm is used to produce a set of $C$ clusters, where $C$ indicates the number of mixture components desired, i.e. the number of individual distributions. The GMM, $\Phi^{av}$, can be described by,

$$\Phi^{av} = \sum_{c=1}^{C} \gamma_c \phi_c(\boldsymbol{z}) = \sum_{c=1}^{C} \gamma_c \, \mathcal{N}(\mathbf{z}; \mu^c, \boldsymbol{\Sigma}^c), \tag{5.2}$$

where $\gamma_c$ represents the prior probability of the $c$th cluster; and $\phi_c(\boldsymbol{z})$ is the $c$th multivariate Gaussian probability density function (PDF) parameterised by mean vector, $\boldsymbol{\mu}_c$, and covariance matrix, $\boldsymbol{\Sigma}_c$. As the model is produced from joint audiovisual feature vectors, parts of the mean and covariance parameters correspond to the separate feature vector inputs. The mean vector, $\boldsymbol{\mu}_c$, of cluster $c$ can be

written as,

$$\boldsymbol{\mu}_c = [\boldsymbol{\mu}_c^a, \boldsymbol{\mu}_c^v] \tag{5.3}$$

where $\boldsymbol{\mu}_c^a$ and $\boldsymbol{\mu}_c^v$ are the audio and visual vector means, respectively. The covariance matrix, $\boldsymbol{\Sigma}_c$, can be written as,

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \boldsymbol{\Sigma}_c^{aa} & \boldsymbol{\Sigma}_c^{av} \\ \boldsymbol{\Sigma}_c^{va} & \boldsymbol{\Sigma}_c^{vv} \end{bmatrix} \tag{5.4}$$

where $\boldsymbol{\Sigma}_c^{aa}$ is the covariance matrix of the audio feature vectors, $\boldsymbol{\Sigma}_c^{vv}$ is the covariance matrix of the visual feature vectors, and $\boldsymbol{\Sigma}_c^{av}$ and $\boldsymbol{\Sigma}_c^{va}$ are the cross-covariance matrices of the audio and visual feature vectors.

Given an input visual feature vector, $\boldsymbol{v}_i$, and the joint audiovisual GMM, $\Phi^{av}$, the maximum *a posteriori* probability (MAP) estimate of the audio feature estimate, $\hat{\mathbf{a}}_i$, can be produced using,

$$\hat{\boldsymbol{a}}_i = \arg\max_{\boldsymbol{a}} \left[ p\left(\boldsymbol{a}_i | \boldsymbol{v}_i, \Phi^{av}\right) \right], \tag{5.5}$$

which can also be expressed as,

$$\hat{\boldsymbol{a}}_i = \boldsymbol{\mu}_c^a + \boldsymbol{\Sigma}_c^{av} \left(\boldsymbol{\Sigma}_c^{vv}\right)^{-1} \left(\boldsymbol{v}_i - \boldsymbol{\mu}_c^v\right). \tag{5.6}$$

Furthermore, estimates from each of the $C$ cluster components in the GMM can be combined to form a weighted summation according to the *posteriori* probability, $w_c(\mathbf{v}_i)$, of the visual feature vector having come from cluster $c$. Accordingly, the

weighted MAP estimate of the audio feature vector can be written as,

$$\hat{\boldsymbol{a}}_i = \sum_{c=1}^{C} w_c(\mathbf{v}_i) \left[ \boldsymbol{\mu}_c^a + \boldsymbol{\Sigma}_c^{av} \left( \boldsymbol{\Sigma}_c^{vv} \right)^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}_c^v) \right], \tag{5.7}$$

where $w_c(\mathbf{v}_i)$ can be obtained from,

$$w_c(\mathbf{v}_i) = \frac{p\left(\mathbf{v}_i | \phi_c^v\right) \gamma_c}{\sum_{c=1}^{C} p\left(\mathbf{v}_i | \phi_c^v\right) \gamma_c}, \tag{5.8}$$

and where $p(\mathbf{v}_i|\phi_c^v)$ is the marginal distribution of the visual feature vector, $\mathbf{v}_i$ having being produced by the Gaussian component $\phi_c^v$.

To find an optimal number of mixture components to use within the model, evaluations are performed on a test set of the speech corpus. For various numbers of components, $C$, a GMM is built using a joint audiovisual feature training set, with the desired number of clusters. Audio estimates from the test set visual features are obtained through application of Equation 5.7. The error between the estimated audio feature vectors, $\hat{\mathbf{a}}_i$, and the original audio feature vectors, $\mathbf{a}_i$, is recorded for that number of components. Typical error metrics used include the mean squared error, percentage difference error, or objective quality and intelligibility measures of the reconstructed spectral-envelope such as the those discussed in Section 2.4.

### 5.3.2 Deep neural networks

Deep neural networks are an extension to standard artificial neural network architectures, as discussed previously in Section 4.3, that have multiple hidden layers stacked together between an input and output layer. Whereas Gaussian mixture models were the major choice for acoustic modelling in speech recognition systems, deep neural networks become more widely applied circa 2009–2012, with

significantly lower word-error-rates (WER) reported [Hinton et al., 2012]. Currently, deep neural networks, and various other neural network architectures, are state-of-the-art in many areas of speech processing, and other fields such as image processing [Krizhevsky et al., 2012; Ciresan et al., 2010]. The power of deep neural networks lies in their ability to extract progressively higher abstract feature representations. An example deep neural network architecture is shown in Figure 5.9, where three hidden layers are stacked between the input and output layers.

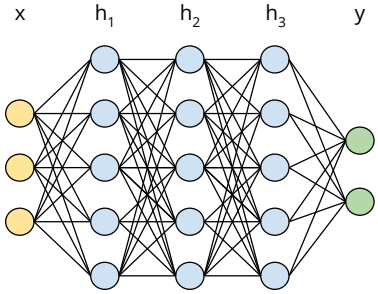



**Figure 5.9:** Standard feed-forward deep neural network architecture with three hidden layers, $h_1$, $h_2$, and $h_3$; between the input layer, $x$, and output layer, $y$. The network is fully-connected, i.e. all units in one layer are connected to all other units in the adjoining layers.

To perform the mapping from the visual to the audio domain, a deep neural network can be expressed simply as,

$$\hat{\mathbf{a}}_i = f(\mathbf{v}_i), \tag{5.9}$$

where $f$ is a feed-forward neural network configured for regression. The function $f$ is comprised of two or more hidden layers between the input and output layers, with the model weight parameters derived from a set of training data using the backpropagation of errors algorithm. Neural networks configured for regression do not have the final application of the softmax function as in Equation 4.5, as is performed for the voicing classification tasks presented in Chapter 4. Rather, the output is linear, which is required for estimating the real-valued audio coefficients.

To derive the required weight parameters for each of the layer connections, the backpropagation of errors algorithm, using gradient descent optimisation, is applied to minimise the mean squared error between the audio features estimated by the network, $\hat{\mathbf{a}}_i$, and the original audio features, $\mathbf{a}_i$. The mean squared error function to minimise is expressed as,

$$E(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{\mathbf{a}}_i - \mathbf{a}_i|_2^2 \,, \tag{5.10}$$

where $N$ is the total number of training samples. In the case of mini-batch stochastic gradient descent, $N$ is equal to the batch size. The weight values are initialised with uniformly distributed random variables in the range $-0.01$ to 0.01. No pre-training is conducted as whilst it is beneficial for preventing overfitting on smaller datasets, it is not so important when training on more, well-balanced data [LeCun et al., 2015]. Although GRID is a comparatively small dataset, it has the benefit of having numerous examples in the data of each word in the grammar.

Aside from obtaining the weight parameters through training of the network, a number of model hyper-parameters can be optimised to further improve the performance of the network. Random or grid search can be used over the set of hyper-parameters to acquire a combination that gives the best performance. Random search tends to reach comparable solutions to grid search quicker within a much shorter search time [Bergstra and Bengio, 2012]. More details of the DNN architecture used are given in Appendix B.3.

## 5.4 Speech reconstruction

Initial experiments are conducted to objectively determine how accurately audio features (LPC and Mel-filterbank) can be estimated from visual features (2D-DCT

and AAM) using the two domain mapping models. Using the results collected in Section 5.2 on spectral smoothing, the accuracy of estimating audio features with various levels of smoothing applied is measured. Furthermore, the two best performing configurations of mapping model, and audio and visual features, are evaluated further with subjective listening tests, to establish whether intelligible audio speech can indeed be reconstructed from visual speech information. The subjective tests are conducted using utterances reconstructed for the female speaker (speaker four in the GRID corpus). Finally, a detailed analysis is presented to determine the characteristics of reconstructed audio speech in the visual-to-audio system as described by Figure 5.1.

### 5.4.1 Objective results

To determine how accurately the audio feature vectors have been estimated, the correlation, $r$, is calculated between the original audio feature vectors, $\mathbf{a}_i$, and the estimates, $\hat{\mathbf{a}}_i$, using,

$$r = \frac{\sum_{i=1}^{n}(\mathbf{a}_i - \bar{\mathbf{a}})(\hat{\mathbf{a}}_i - \bar{\hat{\mathbf{a}}})}{\sqrt{\sum_{i=1}^{n}(\mathbf{a}_i - \bar{\mathbf{a}})^2 \sum_{i=1}^{n}(\hat{\mathbf{a}}_i - \bar{\hat{\mathbf{a}}})^2}}, \tag{5.11}$$

where $\bar{\mathbf{a}}$ and $\bar{\hat{\mathbf{a}}}$ are the means of the original and estimated audio feature vectors respectively. Mean squared error comparisons, whilst informative when considering each audio feature individually, were found to not work well when comparing between the two audio feature representations due to differences in the magnitudes of the errors.

Correlations are reported in Table 5.2 and Table 5.3 for each of the visual feature representations and model combinations, and for the LPC and Mel-filterbank audio features. It can be seen that Gaussian mixture model are superior at estimating

LPC audio features over DNNs, as higher correlations results. Although only a small difference is exhibited, the AAM visual features perform marginally better than 2D-DCT features. For the Mel-filterbank audio features, the DNN achieves slightly higher correlations over using GMMs, with, again, a small tendency for the AAM visual features to outperform the 2D-DCT features.

**Table 5.2:** Correlation scores, $r$, for LPC configurations.

| | DNN | | GMM | |
|---|---|---|---|---|
| Num. coeffs | AAM | 2D-DCT | AAM | 2D-DCT |
| 2 | 0.59 | 0.62 | 0.73 | 0.72 |
| 4 | 0.61 | 0.62 | 0.72 | 0.71 |
| 6 | 0.57 | 0.59 | 0.72 | 0.72 |
| 8 | 0.62 | 0.65 | **0.73** | 0.71 |
| 14 | 0.52 | 0.53 | 0.71 | 0.72 |

**Table 5.3:** Correlation scores, $r$, for Mel-filterbank configurations.

| | DNN | | GMM | |
|---|---|---|---|---|
| Num. channels | AAM | 2D-DCT | AAM | 2D-DCT |
| 4 | 0.82 | 0.81 | 0.79 | 0.81 |
| 7 | 0.83 | 0.82 | 0.79 | 0.81 |
| 10 | 0.83 | 0.82 | 0.81 | 0.81 |
| 15 | 0.83 | **0.82** | 0.81 | 0.81 |
| 20 | 0.82 | 0.82 | 0.81 | 0.81 |

Furthermore, regarding the ability of the models to estimate fewer audio coefficients with greater accuracy, the results show that there is not a great difference between the audio feature estimates with various levels of smoothing applied. For each of the feature and model combinations, except when using a DNN for estimating LPC audio features, there is little difference between the correlations. Accordingly, this suggests that the dimensionality of the audio features should be chosen such that audio speech reproductions have the highest intelligibility possi-

ble, i.e. using LPC features with order $P \geq 8$, and Mel-filterbank channel numbers of $K \geq 15$.

The models selected for the subjective tests are highlighted in bold, and have been chosen as they result in approximately the highest correlations, and have configurations such that there is no overlap between the different types of audio feature, visual feature, and statistical model.

### 5.4.2 Subjective results

The aim of the subjective intelligibility experiments is threefold. First, to examine whether reconstructing audio speech from visual features can produce intelligible speech. Second, to compare the intelligibility of the reconstructed audio with the intelligibility from just the video of the speaker, i.e. lip reading. Third, to examine whether combining reconstructed audio with the video improves intelligibility. To address these questions the subjects are presented with samples from three different multimedia configurations: the reconstructed audio-only, the original video-only, and the reconstructed audio combined with the original video (audiovisual).

To generate the reconstructed audio, four different configurations are examined. Two methods of estimating the time-frequency surface are used, one using a GMM with AAM and 8th-order LPC audio features, and the second using a DNN with 2D-DCT and 15-channel Mel-filterbank audio features. It is apparent that these two systems represent only a small subset of the configurations analysed in Section 5.4.1. However, whilst the results indicate that all of the systems would have been good contenders, it would have been prohibitive to include all combinations (or indeed more combinations) in the subjective experiments as the listening tests would have been too long for the subjects. Instead, the approach chosen was to use two very different configurations to examine their impact on intelligibility. Addition-

**Table 5.4:** Methods of reconstructing speech from visual features.

| Method | Time-frequency surface | Excitation |
|---|---|---|
| GMM_ORIG | GMM + AAM + LPC | Original |
| GMM_UNV | GMM + AAM + LPC | Unvoiced |
| DNN_ORIG | DNN + 2D-DCT + Filterbank | Original |
| DNN_UNV | DNN + 2D-DCT + Filterbank | Unvoiced |

ally, one of the motivating factors for running this set of experiments was to test the hypothesis that intelligible audio speech could in fact be reconstructed from visual speech. These were combined with two methods for creating the speech excitation—using the original voicing and fundamental frequency, and using fully unvoiced excitation. Again, it would be prohibitive to try all combinations of excitation in the listening tests, so preliminary tests determined that the unvoiced excitation gave the most intelligible audio of the three methods introduced in Section 4.2. These two choices of excitation allow the impact of having no knowledge of the voicing/fundamental frequency to be compared to having full knowledge. The four methods are summarised in Table 5.4.

Twenty listeners took part in the tests, which were conducted in a quiet environment with subjects using headphones and positioned in front of a monitor. Each subject was played (in a random order to remove any bias) 12 audio-only sentences, 12 audiovisual sentences, and 3 video-only sentences. The 12 audio sentences, and 12 audiovisual sentences, comprise 3 examples from each of the four configurations in Table 5.4. Only 3 video-only sentences were included as, with no audio present, repeating for the four configurations in Table 5.4 was not required. This gave a total of 27 sentences, all with different utterances. Each listener was allowed to replay the audio/video as many times as they wished before entering the words they heard. The tests were conducted in this manner as a potential application of the work would be to transcribe speech from recordings where lis-

teners would be able to replay the media multiple times. Overall accuracy was calculated by dividing the total number of words correctly identified by the total number of words presented.

Table 5.5 shows the intelligibility (word accuracy) for the four different methods of reconstructing audio, listed in Table 5.4, and are shown for both the reconstructed audio only and when combined with the original video (audiovisual). The intelligibility obtained using the video alone was 50 %, with a range of scores from 0 % to 72 %. This variability is observed in the literature [Summerfield, 1992; Lan et al., 2012], were there exists large variation in lip-reading performance. To control for this, a sufficient number of listeners were required to conducted the experiments, and the scores of all the listeners are averaged. Furthermore, presenting the questions to the listeners as closed-response (using multiple-choice) makes the task easier than if the questions were open-response.

For the GRID grammar shown in Table A.1, the intelligibility that would be expected by chance alone is 19 %, and can be calculated for a single utterance using,

$$(0.25 + 0.25 + 0.25 + 0.1 + 0.04 + 0.25) \div 6 = 0.19, \tag{5.12}$$

where the probability of selecting the correct choice at random for each of *command*, *colour*, *proposition*, and *adverb* is $\frac{1}{4}$; for *digit* it is $\frac{1}{10}$; and for *letter* the probability is $\frac{1}{25}$.

The results show that for both configurations without any prior knowledge of the original excitation (GMM_UNV and DNN_UNV), audio speech can be reconstructed from visual features with intelligibility greater than chance. When these are supplemented by the original video signal, the intelligibility increases further. Audiovisual intelligibility with GMM_UNV is higher than using only the video,

**Table 5.5:** Intelligibility (word accuracy in per cent) and standard error of reconstructed audio-only and audiovisual speech.

| Method | Audio-only (%) | Audiovisual (%) |
|---|---|---|
| GMM_ORIG | 49.2 (5.7) | 60.8 (5.0) |
| GMM_UNV | 40.3 (4.9) | 52.2 (5.5) |
| DNN_ORIG | 37.5 (4.8) | 56.4 (5.7) |
| DNN_UNV | 27.2 (5.3) | 45.8 (4.7) |

and agrees with studies that show that an audiovisual signal is more intelligible over using a single modality [Summerfield, 1992]. The audiovisual intelligibility of DNN_UNV remains lower than visual-only, and is attributed to the lower intelligibility of the audio, at approximately 13 % lower than with GMM_UNV-based audio. Identifying the reason for this difference is not straightforward as the two configurations differ in their audio and visual features as well as the method of estimation. However, informal listening tests indicate that speech produced by a range of different configurations suggests that the audio feature is most important when considering intelligibility, rather than the visual feature or method of estimation. The spectral-envelope produced from estimated LPC coefficients is closer to the original spectral-envelope than that produced by the Mel-filterbank features due to its relative coarseness.

Reconstructing audio using the fundamental frequency and voicing estimated from the original speech gives an absolute increase in intelligibility of 7.5 % over using a purely unvoiced excitation. This demonstrates the importance of voicing and is attributed to several of the vocabulary items requiring voicing to be classified correctly, such as /s/ and /z/ confusions.

Interestingly, the correlation scores from Table 5.2 and Table 5.3 suggest that the DNN_ORIG and DNN_UNV systems would outperform the equivalent GMM systems, with audio feature correlations of 0.82 and 0.73 for the DNN and GMM,

respectively. However, the subjective intelligibility results show that the GMM utterances were more intelligible than the equivalent DNN utterances for all combinations of $f_0$ contour and media. This suggests that correlation between original and estimated audio feature representations is not a good measure of intelligibility for reconstructed utterances. Accordingly, the mean squared error is used for the remainder of this work when evaluating the accuracy of estimated audio features as the results obtained are more informative.

### 5.4.3 Utterance analysis

To further evaluate the thesis of being able to reproduce intelligible audio speech using only visual information with regression models, an analysis is performed on whole utterances and also individual frames of speech in an attempt to determine the characteristics of the reproduced audio speech.

Wideband spectrograms are presented in Figure 5.10 for an original utterance and two reproductions, the first using a GMM with LPC and AAM features, and the second using a DNN with Mel-filterbank and 2D-DCT features. It can be seen in the spectrogram of the original utterance that the formants are clear and that greater spectral detail is present, whereas it is not as readily apparent in the two audio reconstructions. The utterance reproduced from the GMM model appears to be more faithful to the original utterance in comparison to the DNN model, where, although an amount of smoothing is exhibited, there is nevertheless some spectral detail present. This effect is confirmed through informal listening tests, where speech reproductions from LPC coefficient spectral-envelope representations sound more "speech-like" than the utterances derived from the estimated Mel-filterbank features. From the spectral structure of the utterances, it is apparent that the DNN system produces a more sparse spectral structure compared to the GMM

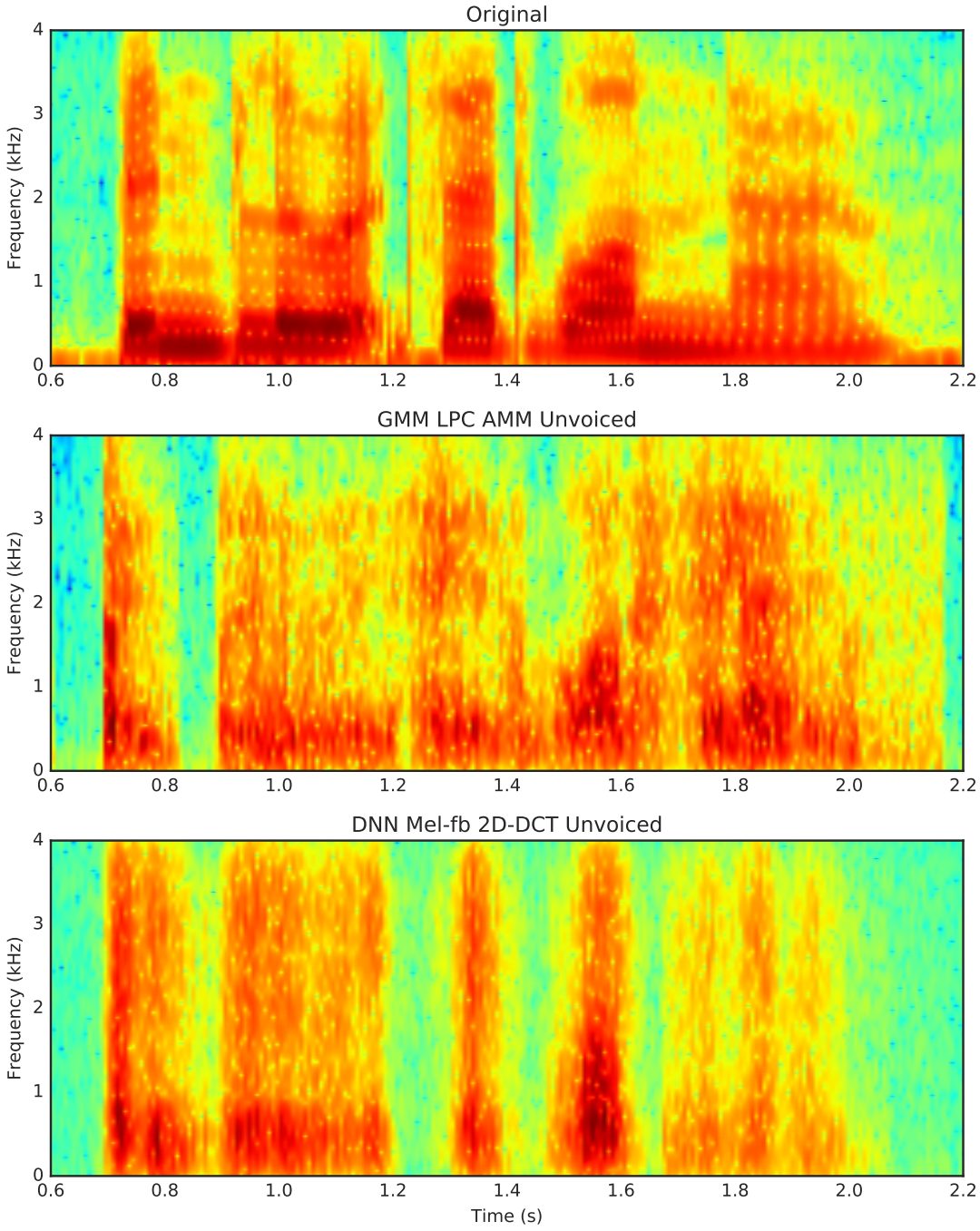



**Figure 5.10:** Wideband spectrograms for the original utterance "bin blue at Z 1 now" spoken by the female speaker, and of reproductions from the GMM and DNN visual-to-audio domain mapping models. Some higher-resolution formant detail is present in the GMM audio reproduction, whereas very little is present in speech reproduced from the DNN.

system, with more energy in the lower frequencies. These observations explain the accuracies achieved in the subjective listening tests, where the GMM system results in greater intelligibility over the DNN system.

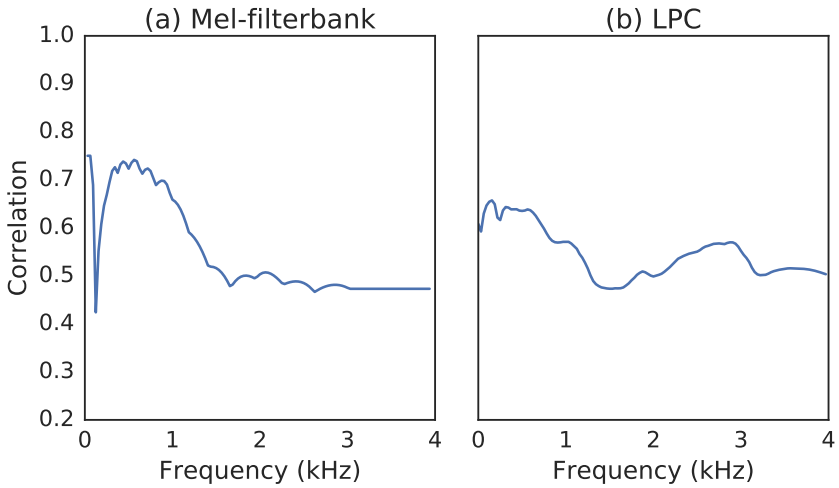



**Figure 5.11:** Correlations of frequency bins between the original and estimated spectral-envelope surfaces for the Mel-filterbank and LPC audio features, respectively.

To further understand the difference in spectral detail, investigations are conducted on the correlations over all test utterances between the original spectral-envelope surface and those estimated by both systems. The difference between the two configurations is shown further in Figure 5.11a for Mel-filterbank, and Figure 5.11b for LPC, where correlations between frequency bins of the original and estimated spectral-envelope surface have been recorded. The DNN system using Mel-filterbank features exhibit strong correlation in the frequencies below 1.8 kHz, and then weaker correlations for frequencies above. Whereas for the GMM system using LPC features, the correlations appear to be somewhat more uniform over the frequency domain, with troughs at 1.5 kHz and 3.1 kHz. The correlations over the frequency domain for the spectral-envelopes confirms what is observed in the spectrograms of the different configuration utterances shown in Figure 5.10, where

the lower-frequencies are better estimated using Mel-filterbank features, and the higher-frequencies are more accurate when using LPC.

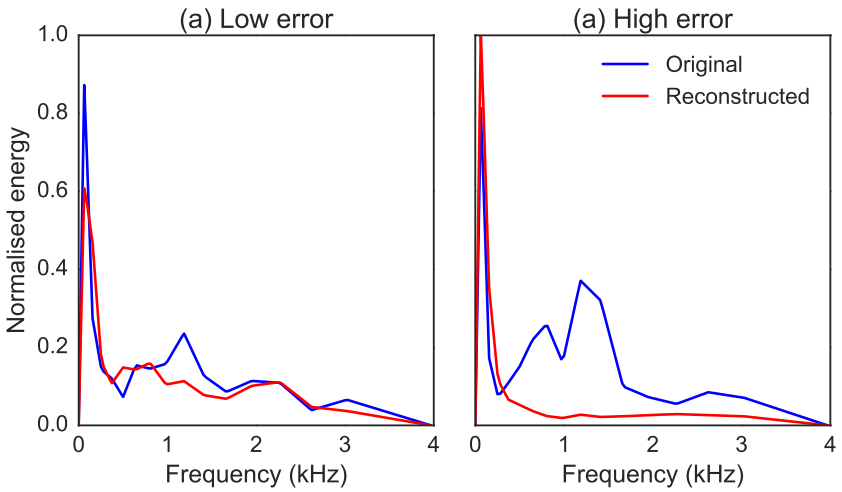



**Figure 5.12:** Comparisons of original and estimated Mel-filterbank features with low and high error spectral-envelope reconstructions for a chosen frame. The left graph shows a reconstructed envelope with very little error, in comparison to the right graph.

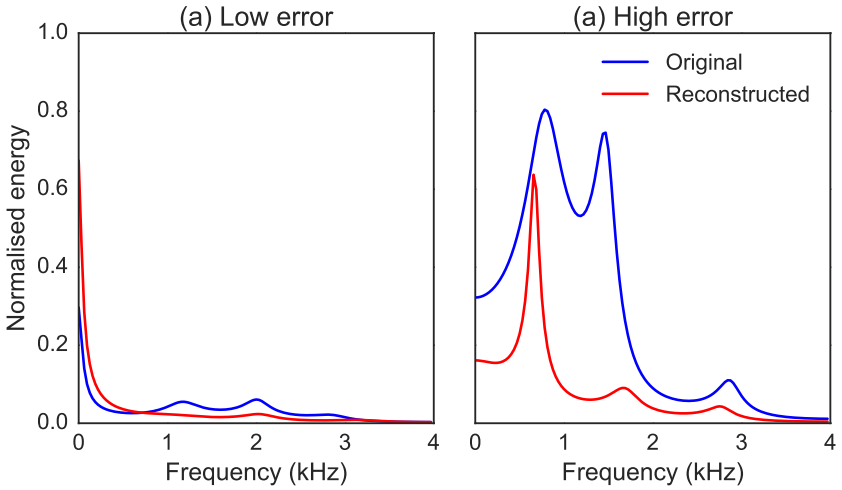



**Figure 5.13:** Comparisons of original and estimated LPC audio features with low and high error spectral-envelope reconstructions for a chosen frame. The left graph shows a reconstructed envelope with very little error, in comparison to the right graph.

Next investigations are conducted at the frame level, where examples of reconstructed spectral-envelopes with low and high-error, when compared to the original spectral-envelopes, are shown in Figure 5.12 for the Mel-filterbank features estimated by a DNN, and in Figure 5.13 for LPC features estimated by a GMM. For both types of audio feature, when there is low error between the original and estimated spectral-envelopes, the estimated audio features have been estimated from input visual features using the visual-to-audio domain mapping models with sufficiently high accuracy. However, when the error is high, it is readily apparent that the spectral-envelopes are significantly dissimilar, causing the resultant speech to have a far lower intelligibility. One issue with using LPC features is that they are not robust to individual errors in the coefficients. That is, errors in the coefficients may lead to the filter becoming unstable [So and Paliwal, 2007], yielding highly-incorrect spectral-envelopes. In comparison, errors in a single channel of the Mel-filterbank features will not introduce errors into other channels, leading to a more stable spectral-envelope representation.

## 5.5 Summary

The experiments conducted in this chapter have shown that it is indeed possible to reconstruct intelligible audio speech signals from visual speech information. In comparison to articulatory speech synthesis models, where information concerning the location of the articulators is available, the information contained within a video of a speaker is limited to that which can be seen of the mouth, with no excitation information obtainable.

The investigations into spectral-smoothing show that speech reproduced from heavily smoothed spectral-envelopes retains high intelligibility in certain configurations, with intelligibility using only a few audio feature coefficients comparable

to that of unprocessed speech. Furthermore, the effect of using the artificial-$f_0$ methods for providing excitation information was explored, with the tendency for the unvoiced and monotone methods to perform best.

The LPC audio feature representations have been shown to outperform Mel-filterbank features for both the spectral smoothing experiments, and for the initial regression visual-to-audio domain mapping models, for both objective scores and subjective intelligibility results. However, one issue with using LPC coefficients is that errors may lead to high instability, where as Mel-filterbanks channel amplitudes do not suffer from this problem. Additionally, despite yielding lower intelligibility scores, the Mel-filterbank features exhibited better correlations at the lower-frequencies over LPC features, where the increase in intelligibility when using LPC features is attributed to the improved correlations at the higher-frequencies.

When considering the mapping models, the Gaussian mixture model system performs best when using LPC audio features, and specifically, when using AAM visual features. For estimating the Mel-filterbank features, the deep neural network system performs best, and specifically when using 2D-DCT visual features. Overall, the highest correlations for the visual-to-audio configurations results from using the DNN to estimate the Mel-filterbank audio features from AAM visual features. If the overall spectral-envelope correlations can be improved, especially in the higher-frequencies, then it is expected that greater intelligibility can be achieved. Accordingly, work presented in the upcoming chapters explores exploiting the power of deep neural network architectures, using Mel-filterbank audio features and AAM visual features, to further improve the results achieved within this chapter.

# Chapter 6

# Classification system

## 6.1 Introduction

In the previous chapter, the use of regression models was explored for performing visual-to-audio domain mapping. That is, input visual features were used to estimate the real-valued and continuous LPC and Mel-filterbank audio feature representations of the spectral-envelopes. A limitation with using regression models for the audio feature estimation is that non-plausible spectral-envelopes can be generated. Whereas errors in the LPC coefficients can cause the filter to become unstable, Mel-filterbank features are more robust to errors in the estimated channel amplitudes, as an error in one channel will not effect the values of other channels. Furthermore, the estimated spectral-envelopes exhibit a large degree of smoothing when compared to the original, and, accordingly, the intelligibility of reconstructed utterances is adversely affected.

The work presented in this chapter explores the idea of applying a clustering-and-classification approach to the task of visual-to-audio domain mapping. The belief here is that by clustering the spectral-envelope information, using vector

quantisation techniques, to produce a codebook, and then estimating the codebook entries, i.e. class labels, using a classification model with input visual features, the problem of estimating non-plausible spectral-envelopes is mitigated. An overview of this approach is shown in Figure 6.1, where deep neural networks are used as the classification model. Furthermore, the quantised audio features, ensuring that similar feature vectors are grouped together, will result in more realistic and reliable spectral-envelopes estimations, with better spectral detail. Thus, the hypothesis is that the intelligibility of the audio speech reconstructions will be greater than the regression system.

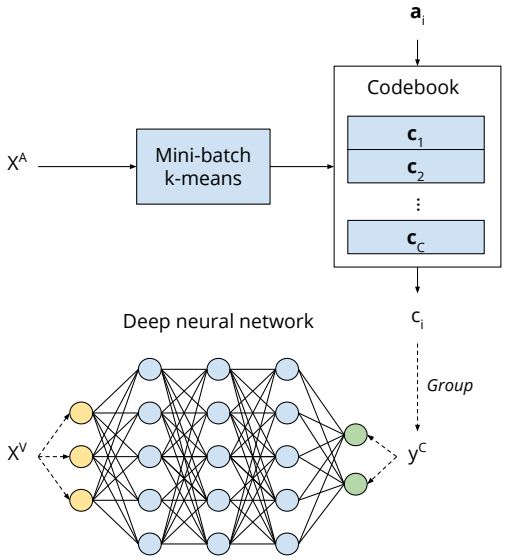



**Figure 6.1:** Overview of proposed system using vector quantisation techniques to produce a codebook of spectral-envelope representations, indexed by a class label. A classification DNN can then be trained using input visual feature vectors and class labels from the associated quantised audio feature vectors.

In addition to the proposed clustering-and-classification framework, the idea of incorporating greater temporal information is explored. The regression system use only static features, i.e. a single visual feature vector is used to predict a single audio feature vector. However, as speech production is a dynamic process, due to, for example, effects of co-articulation and speed of articulator movements,

there is likely a benefit to be had from exploiting longer-range temporal structure. Accordingly, in addition to exploring quantisation of single audio feature vectors in the codebook production stage, the quantisation of grouped audio feature vectors is performed, with estimations from inputs of grouped visual feature vectors.

The remainder of this chapter is organised as follows. In Section 6.2, an overview of codebook production using clustering methods from the area of vector quantisation is given. The classification deep neural network architecture is described in Section 6.3, for estimating the audio feature codebook entry from input visual features. Incorporating longer-range temporal information at the feature-level is described in Section 6.4, for both audio and visual features. Evaluations of the clustering-and-classification methods, using feature-level temporal encoding, proposed in this chapter are presented in Section 6.5. Lastly, a summary of this work is provided in Section 6.6. Subjective intelligibility tests on the audio speech reproductions are presented in Chapter 8.

## 6.2 Vector quantisation

To perform the clustering part of the proposed system, techniques from the area of vector quantisation (VQ) are used. Vector quantisation has application in lossy data compression, such as in video and audio codecs, and is used in numerous areas of speech processing, including voice conversion [Abe et al., 1988] and speech coding [Paliwal and Atal, 1993]. Code-excited linear prediction [Schroeder and Atal, 1985], for example, is a speech coding algorithm that uses both fixed and adaptive codebooks for parametrising excitation information, with LPC coefficients (encoded as LSPs) used to represent the spectral-envelope. The idea behind VQ techniques is that high-dimensional feature vectors can be compressed into a finite set of lower dimension vectors, with a small quantisation error as a result. In

the proposed system, clustering is used to produce a finite codebook of spectral-envelopes, represented by Mel-filterbank channel amplitudes, which can then be converted to spectral-envelopes for speech reconstruction. The aim in performing this step is that class labels can be assigned based on the location of the codebook entry, and then estimated using a classification model.

To produce a codebook, $C$, a clustering algorithm is applied to a set of $N$ training audio feature vectors, $X^A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$. The size of the codebook, i.e. the number of codebook entries, $K$, where $K = |C|$, is chosen such that $K \ll |X^A|$. The set of cluster centres, $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, is found using the mini-batch $k$-means algorithm instead of the classic LBG algorithm [Linde et al., 1980]. The mini-batch $k$-means variant is able, given the large number of training examples, to reach convergence faster with comparable solutions to the standard algorithm [Sculley, 2010].

To initialize the set of cluster centres, $\mathbf{c} \in C$, with size $K$, randomly chosen audio features are selected from the training set, $X^A$. The optimisation problem aims to minimise, over the set of training audio feature examples, the following objective function,

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} ||\mathbf{a}_i - \mathbf{c}_j||^2 \tag{6.1}$$

where $\mathbf{c}_j$ is the mean of the $j$th cluster centre, and $\mathbf{a}_i$ is the $i$th audio feature vector. The index of the cluster centre closest to $\mathbf{a}_i$ can be obtained as follows,

$$j = \arg\min_j ||\mathbf{c}_j - \mathbf{a}_i||^2. \tag{6.2}$$

Gradient descent is employed to minimise the objective function given in Equation 6.1. Mini-batches of size $b$, where $b < N$, of randomly selected training

examples are used to perform updates of the cluster centres for a given number of iterations. The size of the gradient descent taken for each cluster centre is proportional to the number of examples in the mini-batch that have been assigned to that particular cluster. Training is concluded once the total number of processing iterations has been reached. An overview of the clustering process described in this section is given in Figure 6.2, where the codebook is created from the training vectors, $X^A$, and can be used to output a class label for a given input audio feature, $\mathbf{a}_i$.

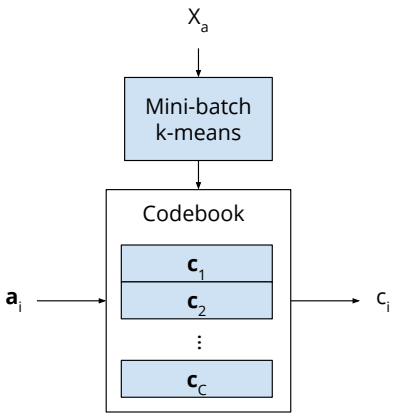



**Figure 6.2:** Overview of the mini-batch $k$-means clustering algorithm applied to a set of audio training features, $X^A$, to produce the codebook, $C$. A class label, $c_i$, can be output by finding the closest cluster centre to $\mathbf{a}_i$.

To evaluate suitable values of $K$, a test set of audio feature vectors are quantised using the codebook, and then compared to the original un-quantised feature vectors to determine the resultant quantisation error. Ultimately, it is desired that audio speech reconstructed from the quantised audio features maintains a level of intelligibility as close as possible to the original utterances. Training audio feature codebooks are produced with sizes $K = \{16, 32, \ldots, 2048, 4096\}$ and the mean squared error, defined in Equation 5.10, and standard deviation, between the original and quantised audio features is recorded.

In Figure 6.3 it can be seen that as the number of cluster centres, $K$, is in-

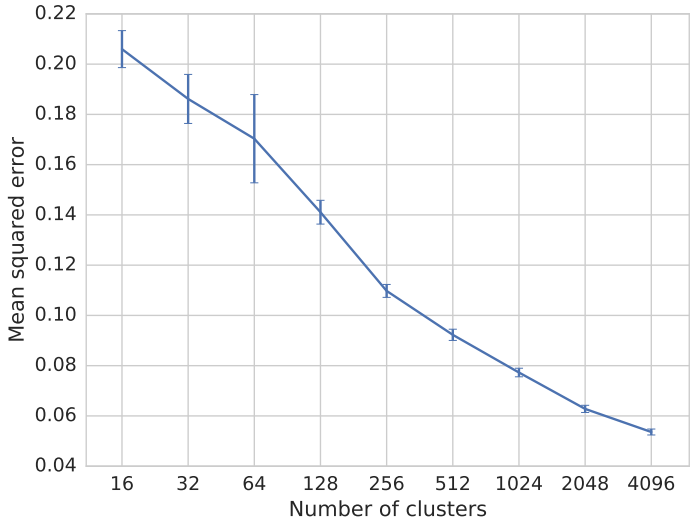



**Figure 6.3:** Mean squared error (with error bars showing a single standard error) between original and quantised audio feature vectors with codebooks of increasing numbers of cluster centres, $K$.

creased, there is a reduction in mean squared error. The error reduces quickly up to $K = 256$ clusters, and then decreases slowly thereafter. The error will continue reducing until it reaches zero when $K = |X^A|$. Spectrograms of audio utterances reconstructed from codebooks with sizes of $K = \{16, 128, 1024\}$, including the original utterance, are shown in Figure 6.4. It is evident that as the size, $K$, of the codebooks increases, greater spectral resolution is retained, with the spectrograms of the original utterance and from audio reconstructed using $K = 1024$ clusters being near identical. With too few clusters there is a distinct loss of spectral-resolution in the higher frequencies, with considerable broadening of the formants exhibited. Informal listening tests, confirming the MSE analysis, indicate that the intelligibility of reconstructed audio utterances is near indeterminable from the original utterances with codebook sizes of $K \geq 512$.

For use within the classification framework, each of the $N$ audio feature vectors

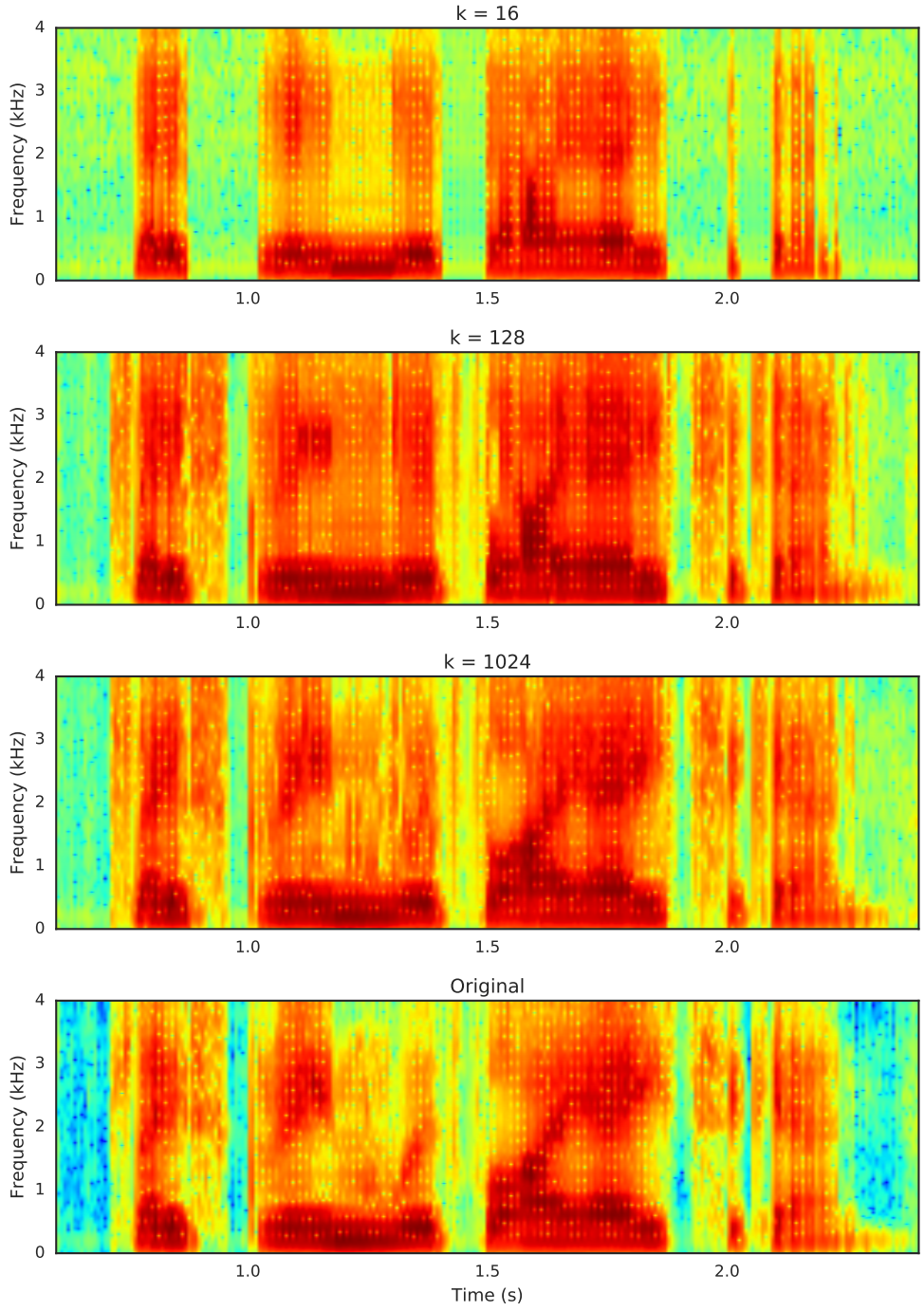



**Figure 6.4:** Spectrograms of the utterance "place green with Y 8 again" spoken by a female speaker, with the original audio and utterances reconstructed from quantised audio features with codebooks of size $K = \{16, 128, 1024\}$.

in the training set, $X^A$, are quantised to give a set of $N$ training class labels, $y^c = \{c_1, \ldots, c_N\}$, derived from the index of the closest cluster centre through application of Equation 6.2. Accordingly, input visual feature vectors, $\mathbf{v}_i$, can be used to estimate the corresponding audio class label, $c_i$, using classification models. In the next section, deep neural networks are explored for performing this classification.

## 6.3 Classification using DNNs

Given input visual feature vectors and a codebook of spectral-envelopes trained as in the previous section, which can be used for assigning class labels to audio features, a classification deep neural network model can be constructed. Deep neural networks configured for classification have application in speech recognition systems where they can be used instead of GMMs for acoustic modelling, and are used to produce posterior probabilities over hidden Markov model (HMM) states [Hinton et al., 2012]. A general overview of deep neural networks for classification is given in Section 4.3, and for regression in Section 5.3.2.

To construct a DNN for performing the visual-to-audio domain mapping, a training set of $N$ visual feature vectors, $\mathbf{X}^V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, and corresponding audio codebook entry labels, $y^c = \{c_1, \ldots, c_N\}$, are required. Weight parameters in the network are uniformly initialised, and mini-batches of training visual features are fed through the network. At the output softmax layer, the categorical cross-entropy loss function (see Equation 4.8) is applied to produce a training error between the estimated class probabilities and the correct class labels. This error is then used to optimise the weights of the network using backpropagation of errors in conjunction with the gradient descent algorithm. As with the regression system, a random search can be performed over the set of model hyper-parameters

to determine an optimal set. More details of the architecture used are given in Appendix B.4.

To obtain an audio feature estimate, $\hat{\mathbf{a}}_i$, a visual feature vector, $\mathbf{v}_i$, is fed through the DNN to produce $K$ class probabilities, with the top class, $\hat{c}_i$, returned using Equation 4.7. The cluster centre indexed by $\hat{c}_i$ can then be output from the audio feature codebook, $C$, to give $\hat{\mathbf{a}}_i$. Interpolation, as per Equation 3.11, can then be applied to the Mel-filterbank feature to obtain a spectral-envelope representation, $X(f, i)$, as required by the STRAIGHT speech reconstruction model.

## 6.4 Feature-level temporal encoding

Thus far, the methods detailed in this chapter, and those presented in the previous chapter on regression, have focused on estimating a single-frame audio feature vector given a single-frame visual feature vector. Velocity and acceleration temporal derivatives can be appended to static feature vectors to introduce some degree of temporal information, however, the context window is typically only on the order of a few frames in width, covering several tens of milliseconds of speech. It is widely recorded in the literature that context plays a important role in speech processing due to phenomena such as co-articulation. Incorporating longer windows of speech information is motivated by psychoacoustic studies of the peripheral human auditory system where it has been suggested that time spans of several hundred milliseconds of speech are integrated, as opposed to the short duration frames most commonly used in speech processing [Sharma et al., 2000]. Additionally, ASR systems have shown benefits from incorporating temporal windows of speech up to 1000 ms in length [Chen et al., 2004], in addition to techniques using shorter frames. In this section, the previous static systems are extended by evaluating different methods of incorporating temporal information at the feature-

level. To complement this, the work in Chapter 7 explores incorporating temporal information at the model-level.

### 6.4.1 Feature-level vector windows

Instead of a static system whereby a single audio feature vector is estimated from a single visual vector at each time instance, windows of feature vectors can be grouped together, to produce audio and visual feature matrices, $\mathbf{A}_i$ and $\mathbf{V}_i$, that comprise windows of size $S^A$ and $S^V$, for audio and visual, respectively. The windows contain an odd number of feature vectors that are centred on a middle vector defined as,

$$\mathbf{A}_i = [\mathbf{a}_{i-w^A}; \ldots; \mathbf{a}_i; \ldots; \mathbf{a}_{i+w^A}], \tag{6.3}$$

$$\mathbf{V}_i = [\mathbf{v}_{i-w^V}; \ldots; \mathbf{v}_i; \ldots; \mathbf{v}_{i+w^V}], \tag{6.4}$$

where $w^A = \frac{S^A-1}{2}$ and $w^V = \frac{S^V-1}{2}$, and the semi-colon operator is used to indicate concatenation of vectors. Larger window widths include greater levels of temporal information. As was performed in Section 6.2 for static vectors, the mini-batch $k$-means algorithm can be applied to the audio feature matrices to produce a codebook where each entry now represents a sequence of $S^A$ static vectors. Accordingly, class labels can be assigned to the windowed feature vectors for use in the classification model.

### 6.4.2 Audio quantisation analysis

One problem with estimating windows of audio feature vectors as opposed to static feature vectors, is that the windowed features need to be manipulated, or

processed, in some capacity so as to result in only a single feature vector for each frame, $i$. This processing is necessary to produce the time-frequency spectral-envelope surface as required by STRAIGHT. Accordingly, three methods for solving this problem are proposed, with intuition for the methods shown in Figure 6.5.

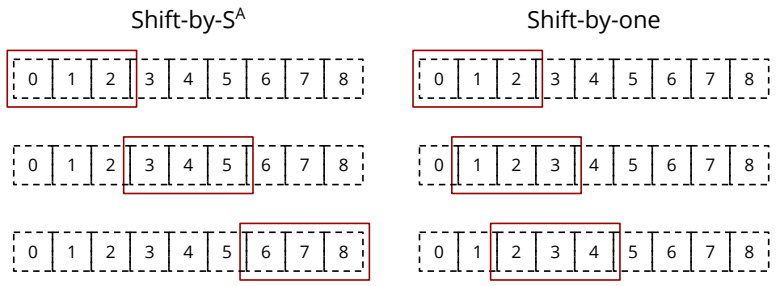



**Figure 6.5:** Intuition for shift-by-$S^A$ (window size) and shift-by-one sliding window techniques for incorporating feature-level temporal information.

1. Shift-by-one: the visual window is shifted forward by a single vector at each time instance and the middle vector in the estimated audio feature matrix, $\mathbf{A}_i$, is selected as the output.

2. Shift-by-$S^A$: the visual window is shifted forward by the size of the audio window, $S^A$, such that output audio matrices are non-overlapping.

3. Overlap-and-add: the visual window is shifted forward by a single vector at each time instance and an audio matrix, $\mathbf{A}_i$, is output. These matrices are then time-aligned and overlap-and-add is applied to form the output vector sequence. As the mean of overlapping windowed features is taken as the output, an element of smoothing to the audio features is introduced.

To evaluate the performance of the three proposed techniques, an audio-only investigation is conducted between the original un-quantised audio feature vectors and the quantised windowed audio feature vectors with the methods applied. Varying feature-level window sizes of $S^A = \{3, 11, 19, 31\}$ are evaluated, with codebook

sizes of $K = \{16, 32, \ldots, 2048, 4096\}$, to see how the error is affected by both codebook size as well as the width of the windowed audio signal. For reference, a single audio frame covers 20 ms of audio signal, with an interval of 10 ms overlap applied. Therefore, for example, a windowed audio feature vector with size $S^A = 11$ will cover 120 ms of audio speech signal.

From Figure 6.6 it can be seen that method three, the overlap-and-add technique, gives the lowest mean squared error across all codebook sizes and audio feature vector window sizes. Methods one and two both show similar trends, converging to roughly the same values for each of the configurations. This suggests that the smoothing effect for method three, by taking the mean of the overlapping audio feature vectors, is having a beneficial effect with regards to lowering the MSE, as opposed to the other methods which have no overlapping output audio features. Furthermore, the method exhibits less erratic trajectories between neighbouring frames leading to lower overall errors.

Additionally, the MSE scores for Method 3 converge to values close to those exhibited in Figure 6.3, for the audio-only evaluations performed on the static features. This suggests that using similar codebook sizes, $K$, for quantising the higher-dimensionality windowed audio feature vectors is sufficient, and not having an adverse effect on the reconstructed audio feature vectors.

### 6.4.3 Visual-to-audio evaluation

Visual-to-audio mapping experiments are conducted to evaluate the effect of estimating the codebook class labels of the quantised windowed audio feature vectors from a classification deep neural network given input windowed visual feature vectors, with the application of method 3 to reconstruct the audio feature vectors for each frame.

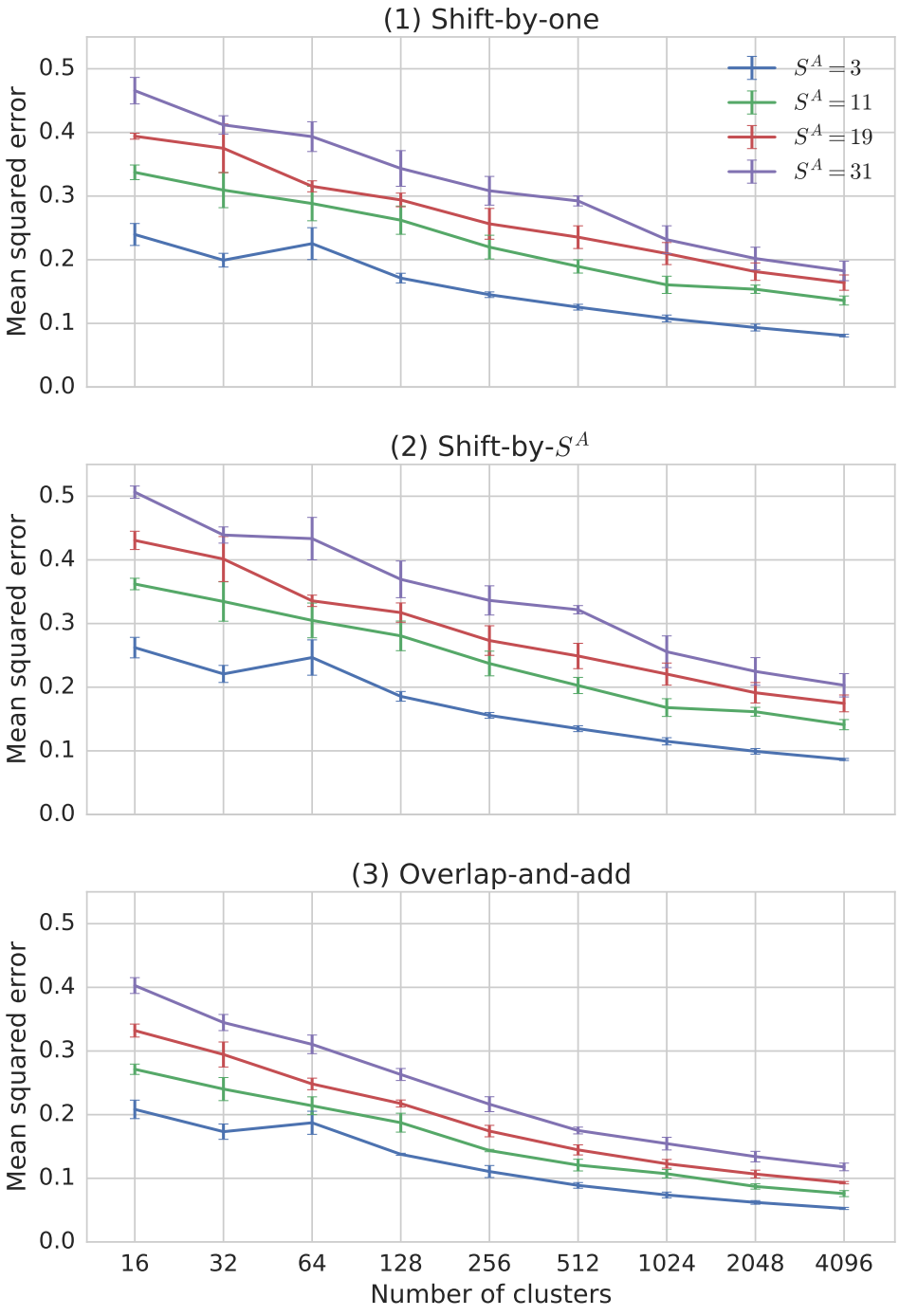



**Figure 6.6:** Mean squared errors (and error bars showing a single standard error)for the three sliding window techniques, for feature-level temporal encoding, between the original audio feature vectors and their quantised versions.

Numerous classification deep neural network systems are trained using a different codebook model size, $K$, visual sliding window size, $S^V$, and audio sliding window size, $S^A$. The window sizes for both the audio and visual features are selected from the set $\{3, 7, \ldots, 31, 35\}$. For all combinations of $S^A$ and $S^V$, the visual-to-audio mapping models are evaluated with varying windowed audio feature codebook sizes of $k = \{512, 1024, 2048, 4096\}$. The MSEs that are reported in Table 6.1 are for the model that yields the lowest error across the four codebook sizes. That is, the output error is shown for only the single best-performing system with codebook size $K$ for each combination of $S^A$ and $S^V$. The errors recorded were for utterances from the female speaker.

**Table 6.1:** Static mean squared error of the estimated audio from a deep neural network and the original audio with varying audio and visual sliding window sizes.

| $S^A$ \ $S^V$ | 3 | 7 | 11 | 15 | 19 | 23 | 27 | 31 | 35 |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0.483 | 0.431 | 0.419 | 0.411 | 0.403 | 0.399 | 0.394 | 0.397 | 0.396 |
| 7 | 0.417 | 0.383 | 0.372 | 0.364 | 0.357 | 0.354 | 0.354 | 0.355 | 0.348 |
| 11 | 0.390 | 0.363 | 0.353 | 0.345 | 0.341 | 0.337 | 0.337 | 0.336 | 0.336 |
| 15 | 0.383 | 0.356 | 0.344 | 0.337 | 0.337 | 0.336 | 0.331 | 0.331 | 0.332 |
| 19 | 0.379 | 0.351 | 0.341 | 0.332 | 0.332 | 0.328 | 0.326 | 0.327 | 0.326 |
| 23 | 0.380 | 0.346 | 0.335 | 0.329 | 0.326 | 0.324 | 0.325 | 0.321 | 0.322 |
| 27 | 0.383 | 0.346 | 0.337 | 0.330 | 0.326 | 0.323 | 0.323 | 0.321 | 0.321 |
| 31 | 0.383 | 0.347 | 0.333 | 0.325 | 0.322 | 0.321 | 0.320 | **0.318** | **0.318** |
| 35 | 0.386 | 0.354 | 0.337 | 0.332 | 0.326 | 0.322 | 0.319 | **0.318** | 0.320 |

From Table 6.1, it can be seen that the general trend is for the mean squared error to decrease as the window sizes for both the audio and visual features increase. As the window size increases from 3 to 23 frames, the MSE decreases quickly, after which it remains relatively constant. The lowest MSE is recorded for a visual window size of $S^V = 31$ and an audio window size of $S^A = 31$. That is, using

320 ms of visual information to estimate 320 ms of audio information; and for reference, with a codebook size of $K = 2048$. This is in comparison to a static system whereby 20 ms of audio is estimated from an equivalent length of visual information.

In the next section, objective intelligibility evaluations are performed for a selection of the best performing feature-level configurations for both a male and female speaker.

## 6.5 Objective intelligibility evaluation

To evaluate the improvement in intelligibility of speech reconstructions using the clustering-and-classification framework proposed in this chapter, visual-to-audio reconstruction experiments are conducted with objective intelligibility evaluations performed. In particular, the accuracy of audio feature estimations given the various feature-level methods and window sizes is explored, with objective intelligibility measures applied to reconstructed utterances to establish configurations with which to perform further subjective listening tests in Chapter 8. Furthermore, a more detailed analysis of the reconstructed utterances is performed to gain a better understanding of the behaviour of the method proposed in this chapter.

### 6.5.1 Objective experiments

To reconstruct time-domain audio utterances, the spectral-envelope information is obtained from the audio features estimated by various system configurations with audio and visual combinations of $S^A = \{23, 27, 31, 35\}$ and $S^V = \{31, 35\}$, respectively. This set of audio and visual window sizes was found to give around the lowest MSE values, as per Table 6.1. To obtain the excitation information, band-

aperiodicity features estimates are obtained from the estimated audio features, $\hat{\mathbf{a}}_i$, and joint codebook, $C^{ap}$, through application of Equation 4.12. Additionally, the monotone artificial-$f_0$ method (from Section 4.2) is used to provide a fundamental frequency contour. The regression models used in these experiments are updated to use AAM visual features, to remain in keeping with the proposed feature-level models. Evaluations of the reconstructed audio utterances, for a male and female speaker, are performed using STOI and PESQ.

**Table 6.2:** STOI intelligibility scores for female speaker with feature-level method.

| $S_A$ \ $S_V$ | 23 | 27 | 31 | 35 |
|---|---|---|---|---|
| 23 | — | — | 0.737 | 0.737 |
| 27 | — | — | 0.737 | 0.737 |
| 31 | 0.737 | 0.737 | **0.740** | 0.739 |
| 35 | 0.739 | 0.740 | 0.739 | 0.738 |

**Table 6.3:** PESQ scores for female speaker with feature-level method.

| $S_A$ \ $S_V$ | 23 | 27 | 31 | 35 |
|---|---|---|---|---|
| 23 | — | — | 1.665 | 1.666 |
| 27 | — | — | 1.656 | 1.656 |
| 31 | 1.653 | 1.652 | 1.668 | **1.671** |
| 35 | 1.660 | 1.664 | 1.666 | 1.665 |

Table 6.2 and Table 6.3, show scores for the female speaker of the STOI and PESQ objective measures, respectively. There is little difference for both metrics across the various system configurations as the scores are within a similar range. As was observed in Table 6.1, the configurations that resulted in the lowest mean squared error scores also resulted in among the highest STOI and PESQ scores. An audio feature and visual feature window size of $S^A = S^V = 31$ gives the best STOI

score of 0.740, and close to the best PESQ score of 1.671. For comparison, speech reconstructed using the static-only method of audio feature estimation (where $S^A = S^V = 1$) had a STOI of 0.507 and PESQ of 0.987, which is significantly lower and indicates the importance of the wide temporal windows. Furthermore, the results from the best configuration compare favourably to the regression system, which achieves a STOI of 0.607 and a PESQ of 1.353.

**Table 6.4:** STOI intelligibility scores for male speaker with feature-level method.

| $S_A$ \ $S_V$ | 23 | 27 | 31 | 35 |
|---|---|---|---|---|
| 23 | — | — | 0.729 | 0.731 |
| 27 | — | — | 0.727 | 0.727 |
| 31 | 0.725 | 0.731 | 0.727 | 0.728 |
| 35 | **0.735** | 0.732 | 0.733 | 0.734 |

**Table 6.5:** PESQ scores for male speaker with feature-level method.

| $S_A$ \ $S_V$ | 23 | 27 | 31 | 35 |
|---|---|---|---|---|
| 23 | — | — | 2.052 | **2.055** |
| 27 | — | — | 2.031 | 2.025 |
| 31 | 2.017 | 2.022 | 2.023 | 2.025 |
| 35 | 2.027 | 2.030 | 2.038 | 2.038 |

Table 6.4 and Table 6.5, show scores for the male speaker of the STOI and PESQ objective measures, respectively. A similar trend is seen for the male speaker as occurs with the female speaker, where the range of scores within each measure are similar. The window sizes for which STOI is maximised occur with $S^A = 35$ and $S^V = 23$ with a score of 0.735, and for PESQ a score of 2.055 is achieved using windows of $S^A = 23$ and $S^V = 35$. In comparison, using static-only estimation, STOI and PESQ scores are significantly lower at 0.390 and 0.974, respectively.

Interestingly, although the STOI scores between the male and female speaker are within a similar range, the male speaker achieves higher PESQ scores with 2.055 for the best configuration in comparison to the female with a score of 1.671. Again, the results compare favourably to the regression system, which, for the male speaker, resulted in a STOI of 0.604 and a PESQ of 1.700.

The improvement in objective intelligibility scores for both speakers using the clustering-and-classification method indicates the superior audio feature estimation performance of this method over the regression system, and demonstrates the benefits of using longer-range temporal information. Next, an analysis of the reconstructed utterances is conducted to elucidate further the increase in objective performance.

### 6.5.2 Utterance analysis

To evaluate further the performance of the clustering-and-classification framework with feature-level temporal encoding proposed in this chapter, an analysis of reconstructed audio utterances is performed. Audio analysis performed on reconstructed utterances in Chapter 5, achieved using wideband spectrograms and correlations of frequency between original and estimated spectral-envelopes, demonstrated that the regression visual-to-audio models were unable to reproduce higher-frequency spectral detail, although the reproduction of the lower-frequencies was more faithful. Accordingly, given the improvement in object intelligibility scores, the methods presented in this section will ideally show more accurate spectral detail as the audio features are selected from a codebook as opposed to being estimated directly.

Wideband spectrograms are shown in Figure 6.7 for an original utterance, and utterances reconstructed using the regression and clustering-and-classification

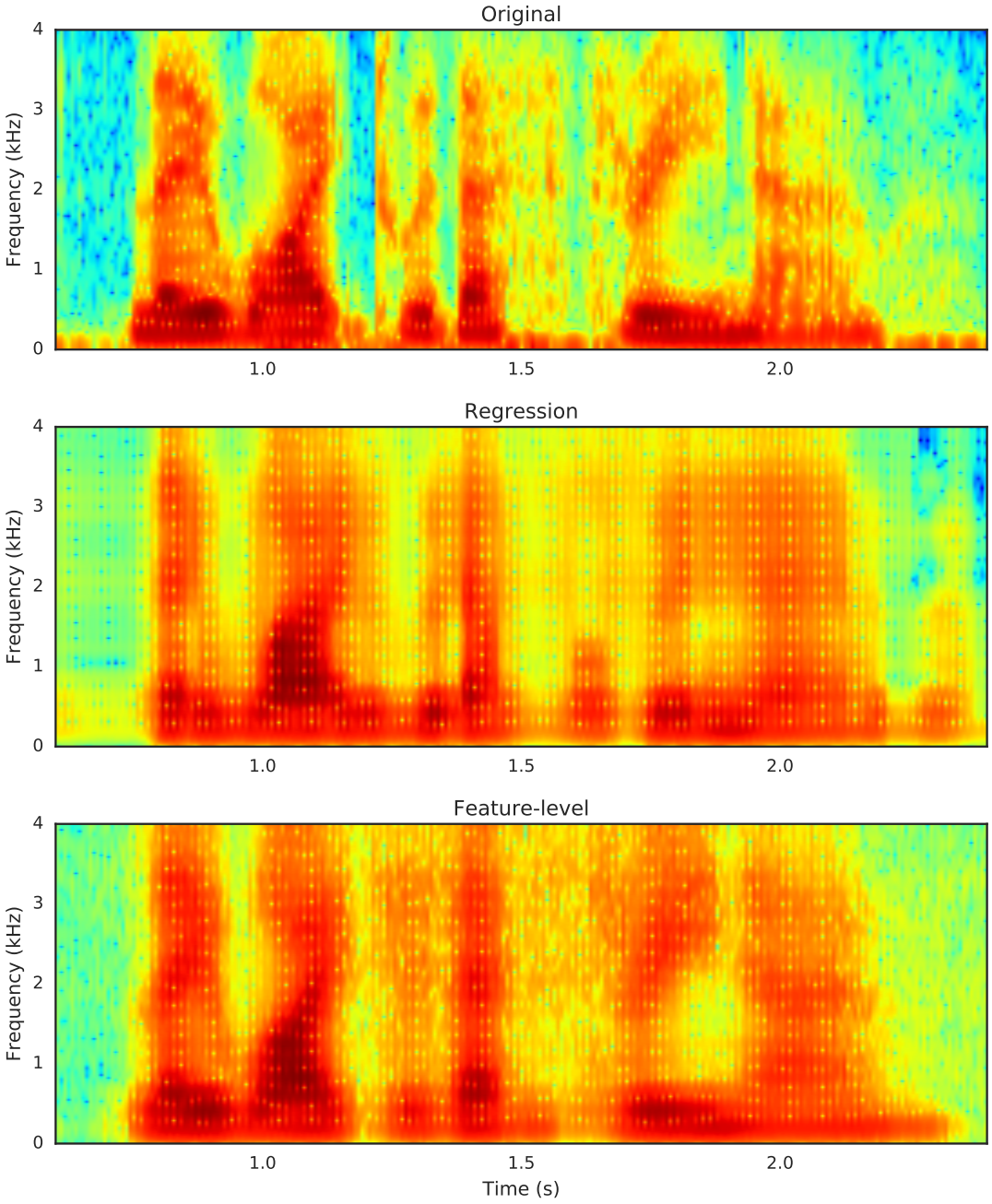



**Figure 6.7:** Wideband spectrograms of the original utterance "lay white with F 3 now" from the female speaker, and reproductions from the regression and feature-level clustering-and-classification visual-to-audio domain mapping models.

models. The regression system, using AAM features, exhibits roughly the same trends as was found for the previous DNN regression system when using 2D-DCT visual features. Although the lower-frequencies have been produced faithfully, there is still little high-frequency spectral detail, with an apparent smoothing noticeable in the general spectral-envelope structure across the utterance. The spectrogram for the feature-level model shows that there is an apparent increase in the higher-frequency spectral detail, with less smoothing found in comparison to the regression system. For example, there is evidence of formants other than the first, $F_1$, around 1.6 and 2 seconds.

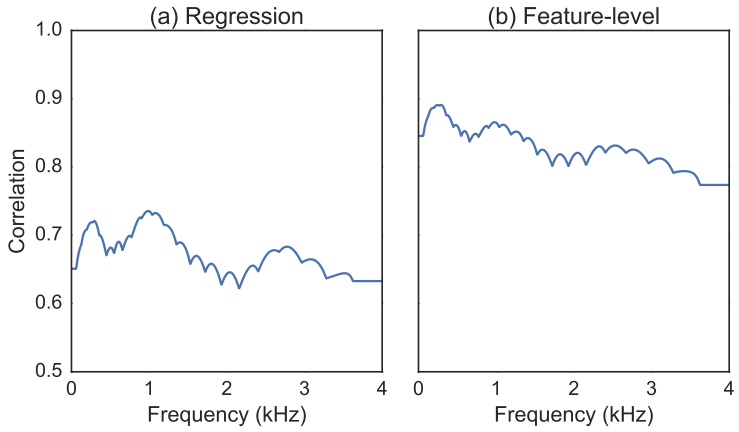



**Figure 6.8:** Correlations of frequency bins between the original and estimated spectral-envelope surfaces for the regression and feature-level models, respectively.

As the error measures give only a single output describing the difference between the original and estimated audio feature vectors, the accuracy across frequency is combined into one value. However, in attempt to further understand the ability of the feature-level methods to produce utterances with higher intelligibility, it is informative to analyse the accuracy of the estimates across frequency. Accordingly, correlations between the original and reconstructed estimated spectral-envelope surfaces are calculated over the set of test utterances. Figure 6.8 shows correlations between the frequency bins of original and reconstructed spectral-envelopes from

the regression and feature-level methods for the female speaker. It can be seen that the regression system exhibits considerably lower correlations across all frequencies in comparison to the feature-level system, which corresponds to the difference in spectral detail observed in the spectrograms in Figure 6.7. For the feature-level system, the correlations are more uniform over the range of frequencies, and are generally very strong with all $r$ values around 0.8 or higher. Interestingly, for both methods, there is a similar trend in the frequency regions exhibiting greater correlation, with visible peaks around 200 Hz, 1.1 kHz, and 2.6 kHz.

The spectrograms and correlation analysis clearly show the superior performance of the clustering-and-classification approach to estimate more accurate spectral-envelope surfaces. The results indicate that estimating codebook entries using classification models, including using longer-range temporal information, is preferable over estimating the audio feature coefficients directly using regression methods.

## 6.6 Summary

The experiments conducted within this chapter have shown that the proposed clustering-and-classification framework to perform the visual-to-audio domain mapping yields considerable improvements in objective intelligibility scores over the regression system. Vector quantisation techniques are applied to produce a codebook of windowed audio features from which class labels can be assigned based on the location of entries within the codebook. The benefit of using the audio codebook, as opposed to directly estimating the audio features using regression, is that more accurate spectral-envelopes can be produced as a result. To estimate the class labels from input visual features, deep neural networks configured for classification are utilised.

To improve further the results using the clustering-and-classification method, investigations are conducted on incorporating longer-range temporal information at the feature-level. Windowed audio features, covering various time periods, are quantised and then assigned class labels which are subsequently estimated from windows of input visual features. For both the male and female speakers, the objective intelligibility measures were maximised when using audio and visual window sizes covering around 300 ms of speech. For obtaining a single spectral-envelope surface at each time instance, three methods are presented, with the overlap-and-add method outperforming the other two. The smoothing introduced using this method yields less erratic trajectories at the audio feature boundaries, and serves to reduce the error between original and estimated audio utterances. The best performing configuration for the female speaker, using $S^A = 31$ and $S^V = 31$, is investigated further in Chapter 8 with subjective listening tests.

In the next chapter, incorporating temporal information is explored at the model-level using two approaches. The idea being that the models will determine the inherent temporal structure exhibited in the input audio and visual signals, as opposed to directly encoding the information at the feature-level using windowed feature vectors.

# Chapter 7

# Model-level features

## 7.1 Introduction

In the previous chapter, a clustering-and-classification framework was proposed for performing the visual-to-audio domain mapping, resulting in significant improvements over the regression system as described in Chapter 5. These improvements were achieved as a consequence of incorporating longer-range temporal information, which allows for phenomena such as co-articulation to be modelled, at the feature-level.

As an alternative approach, the work in this chapter explores encoding temporal information at the model-level, where the inherent temporal structure of the audio and visual signals is modelled. As with the previous regression and clustering-and-classification systems, the aim is to produce audio feature estimates from visual speech information for generating spectral-envelope time-frequency surfaces for input into the STRAIGHT speech production model. To investigate incorporating temporal information at the model-level, two methods are considered: Viterbi decoding and recurrent neural networks.

In the first method, Viterbi decoding is applied to a temporal sequence of vectors consisting of probabilities associated with Mel-filterbank feature codebook entries, where, for each time instance, the probabilities are estimated using a deep neural network from a single visual speech vector. The Viterbi dynamic programming algorithm has application in hidden Markov models for discovering the most likely sequence of hidden states (a path), and associated probability, given a sequence of observed vectors. The technique, as applied to HMMs, has application in various areas of speech processing, including ASR [Rabiner and Juang, 1993] and speech synthesis [Hunt and Black, 1996]. In this work, the algorithm is used to output the sequence of codebook entries for which the path has the highest cumulative probability.

The second approach explores using recurrent neural networks (RNN) with the long-short term memory (LSTM) architecture for estimating sequences of audio codebook entries given corresponding sequences of input visual features. The models have shown successful application in numerous areas of speech processing, and associated fields of deep learning, including ASR [Graves et al., 2013b], object recognition and labelling [Valentini-Botinhao et al., 2011], and TTS [Fan et al., 2014]. The LSTM architecture offers numerous benefits over standard RNN implementations, such as model stability and being able to model considerably longer-range temporal dependencies in the data.

The remainder of this chapter is organised as follows. The application of the Viterbi algorithm to sequences of estimated codebook entry probabilities using first-order transitions is discussed in Section 7.2. In Section 7.3, recurrent neural networks using the long short-term memory architecture are explored, in an attempt to model longer-range temporal dependencies than can be achieved using the Viterbi method. Both methods are evaluated objectively in Section 7.4, with an audio analysis performed on reconstructed utterances. The best performing

method and configuration from this section are evaluated further using subjective listening tests in Chapter 8. Lastly, the work presented in this chapter is summarised in Section 7.5.

## 7.2 Viterbi decoding

In this section, the first of the two model-level approaches is presented, where Viterbi decoding is explored for incorporating temporal information. The Viterbi dynamic programming algorithm, proposed by Viterbi [1967], has application for decoding convolution codes for cellular and other communication protocols, and in hidden Markov models for determining the most likely sequence of hidden states resulting in a sequence of observed outputs. In automatic speech recognition scenarios, HMMs are typically used to model words or sub-word units, where the selected HMM is chosen based on it being the model most likely to have output a sequence of observed audio feature vectors.

In this work, the Viterbi algorithm is applied to a sequence, with size $T$, of codebook entry vectors, with size $K$. The codebook entries each have an associated probability as estimated by a neural network from an input visual feature. These estimated probabilities can be obtained by using the static ($S^A = S^V = 1$) clustering-and-classification model from the previous chapter, where the probabilities of the codebook entries represent the emission probabilities in the Viterbi algorithm, and take the form of posterior probabilities, $P(\mathbf{v}|c_j)$. For use in the Viterbi algorithm, these posterior probabilities need to be converted to class-conditional probabilities, $P(c_j|\mathbf{v})$, which can be achieved using Baye's theorem as follows,

$$P(c_j|\mathbf{v}) \propto \frac{(\mathbf{v}|c_j)}{P(c_j)}, \tag{7.1}$$

where $P(c_j)$ is the class prior probability [Dahl et al., 2012]. The class priors can be obtained from the training data by normalising the frequency counts of each codebook entry label.

The sequence of codebook entries can be thought of as a $K \times T$ emission matrix, $\mathbf{B}$, described by:

$$\boldsymbol{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,T-1} & b_{1,T} \\ b_{2,1} & b_{2,2} & \dots & b_{2,T-1} & b_{2,T} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{K-1,1} & b_{K-1,2} & \dots & b_{K-1,T-1} & b_{K-1,T} \\ b_{K,1} & b_{K,2} & \dots & b_{K,T-1} & b_{K,T} \end{bmatrix}, \tag{7.2}$$

where each column contains the $K$ codebook entry probabilities estimated by a deep neural network at each time instance $t = \{1, \dots, T\}$. By performing an arg max operation over the columns in $B$, the top class estimated at each time instance can be obtained. This would give the set of outcomes as described for the static clustering-and-classification system in Section 6.3. In the static system, a single input visual vector is mapped to a single class label, i.e. the class with highest probability, which is then used to output an audio feature from the codebook. However, the static approach does not attempt to incorporate longer temporal information, which was shown to be beneficial in the previous chapter.

The intention of using the Viterbi algorithm in this work is to determine the most likely sequence of codebook entries which can be used to output a whole sequence of audio feature vectors, $[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_T]$, to derive an entire time-frequency spectral-envelope surface for use in STRAIGHT.

Given matrix $B$, from Equation 7.2, of $K$ codebook entry probabilities esti-

mated by a neural network for a sequence length $T$, the total possible number of class label sequences is $K^T$, which, for even small values of $K$ and $T$, is a tremendously large number. However, a great deal of the class probabilities will be close to, or in fact, zero, suggesting that they are unlikely to be in the most probable sequence. Accordingly, the Viterbi algorithm can be used to determine the sequence of class labels, with the greatest overall probability, by ignoring all class label sequences except the one that is most probable.

To introduce temporal information into the model, and as is required by the Viterbi algorithm, a transition probability matrix, $\mathbf{A}$, is produced. First-order transition probabilities are generated by processing a set of training utterances. Firstly, the audio features for each training utterance are quantised using a static codebook, $C$, (see Equation 6.2), to give class label sequences for each utterance, described by

$$\{c_1, c_2, \ldots, c_{T-1}, c_T\}, \tag{7.3}$$

where $c_t$ is the codebook entry at time $t$, and the number of frames and associated class labels for the utterance equals $T$. Then, for all pairs of contiguous class labels, $c_i$ and $c_{i-1}$, the number of times that a particular class is preceded by another class is recorded. This can be achieved by performing successive updates to the transition matrix using

$$\mathbf{A}_{i,j} \leftarrow \mathbf{A}_{i,j} + 1, \tag{7.4}$$

where the location at $\mathbf{A}_{i,j}$ is increased by one for every occurrence of class $c_i$ preceding class $c_j$. Once all first-order class occurrences have been recorded, they are normalised to convert these frequency counts to probabilities. Accordingly, given a codebook, $C$, with size $K$, a matrix of first-order transition probabilities

is produced with $K$ rows and $K$ columns, described by

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,K-1} & a_{1,K} \\ a_{2,1} & a_{2,2} & \dots & a_{2,K-1} & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{K-1,1} & a_{K-1,2} & \dots & a_{K-1,K-1} & a_{K-1,K} \\ a_{K,1} & a_{K,2} & \dots & a_{K,K-1} & a_{K,K} \end{bmatrix}. \tag{7.5}$$

Furthermore, a set of initial class probabilities, defined as $\pi$, with size $K$, can be obtained by normalising the frequency count of the first class, $c_1$, of each training utterance sequence.

Given the transition probability matrix, $A$, emission matrix, $B$, and initial probabilities, $\pi$, the Viterbi algorithm can be applied to determine, and output, the sequence of codebook entries with greatest cumulative probability. The algorithm is defined as the following recurrence relation

$$\alpha_{k,t} = \max_{\text{over } j} \Big( \alpha_{j,t-1} a_{j,k} \Big) b_{k,t} \text{ for } 1 \leq k \leq K,\ 2 \leq t \leq T, \tag{7.6}$$

where $\alpha_{j,t}$ is the cumulative probability of codebook entry, $j$, after emitting the first $t$ observed vectors and having travelled through the sequence of $t-1$ preceding class labels with the highest probability. The initial output of the recurrence relation, $\alpha_{k,1}$, can be obtained from

$$\alpha_{k,1} = \pi_k b_{k,1}, \tag{7.7}$$

where $\pi_k$ is the initial probability of class $k$, and $b_{k,1}$ is the associated emission probability. Through successive applications of Equation 7.6, the final cumulative

probability is obtained using

$$P(c_1, \ldots, c_T) = \max_{\text{over } j} \Big( \alpha_{j,T} \Big), \tag{7.8}$$

where $P(c_1, \ldots, c_T)$ is the cumulative probability of the best sequence of output codebook entries.

When applying Equation 7.6 to give the probability of the sequences, it is also necessary to store the actual values of the path taken. Accordingly, the sequence of codebook entries with the greatest cumulative probability can then be output. A matrix, $\mathbf{S}$, is defined for storing the most likely path at time $t$, and is updated using

$$\mathbf{S}_{k,t} = \arg\max_{\text{over } j} \Big( \alpha_{j,t-1} a_{j,k} \Big). \tag{7.9}$$

One issue in implementing the Viterbi algorithm using multiplication operations on a large number of probability values, all of which are less than or equal to one, is that some of the values produced will be extremely low, and, accordingly, below the range of floating point numbers. To solve this issue, one approach is to apply the logarithmic transformation to the probabilities and perform addition instead of multiplication. Thus, Equation 7.6 can be re-written as

$$\alpha^L_{j,t} = \gamma \max_{\text{over } j} \Big( \alpha^L_{j,t-1} + a^L_{j,k} \Big) + (1 - \gamma) b^L_{k,t}, \tag{7.10}$$

where the superscript $L$ indicates that the logarithmic transform has been applied. The coefficient, $\gamma$, is incorporated into the equation to allow for assigning a greater weight to either the transition or emission matrices. For example, with $\gamma = 1$ there will be no contribution from the emission matrix, and with $\gamma = 0$ there will be no contribution from the transition matrix. With $\gamma = 0.5$, both matrices have equal

weight.

The application of the Viterbi algorithm for incorporating temporal information at the model-level is investigated in Section 7.4. Evaluations are conducted to determine the optimum weighting coefficient, $\gamma$, and then objective intelligibility measures are applied. In the next chapter, recurrent neural networks are explored for model-level temporal encoding.

## 7.3 Recurrent neural networks

In this section, the second approach for incorporating temporal information at the model level is explored using recurrent neural networks (RNN) using the long short-term memory (LSTM) architecture. Recurrent neural networks have shown successful application for unsegmented handwriting recognition tasks [Graves and Schmidhuber, 2009], end-to-end (predicting phonemes from input audio features) training of speech recognition systems [Graves et al., 2013b], and for speech enhancement [Weninger et al., 2015]. The application of these neural network architectures is investigated for predicting output sequences of audio codebook entries from input sequences of visual feature vectors, to produce the necessary time-frequency spectral-envelope surface as required by STRAIGHT for audio speech reconstructions.

Recurrent neural networks are an extension of standard neural networks and are able to model dynamic processes by, in effect, introducing a feedback loop into the standard architecture. A sequence of $T$ input visual vectors, $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$, is passed through hidden layer weight connections to produce a hidden vector sequence, $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$, and finally the output vector sequence, $\{y_1, \ldots, y_T\}$ which comprises class labels pertaining to audio codebook entries. The hidden vector,

$\mathbf{h}_t$, at each time instance can be obtained through application of

$$\mathbf{h}_t = \sigma\left(\mathbf{W}_{vh}\mathbf{v}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}\right), \tag{7.11}$$

where $\sigma$ is the rectified linear unit activation function, $\mathbf{W}_{vh}$ are the input layer to hidden layer weights, and $\mathbf{W}_{hh}$ are the hidden to hidden layer weights. The bias terms have been omitted for clarity. Element, $y_t$, of the output sequence can be obtained through application of

$$y_t = \mathbf{W}_{hy}\mathbf{h}_t, \tag{7.12}$$

where $\mathbf{W}_{hy}$ are hidden layer to output layer weights. Figure 7.1 shows how the output of the hidden layer is propagated through time to the hidden layers at the subsequent time instances.

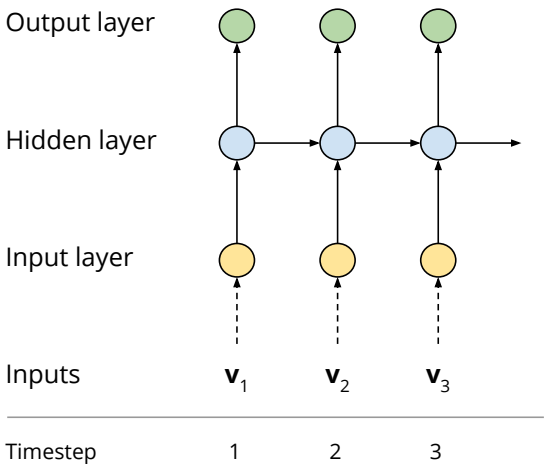



**Figure 7.1:** Outputs from recurrent neural networks are a function of the current input vector and of the previous hidden layer outputs.

### 7.3.1 Long short-term memory architecture

In the LSTM architecture, proposed by Hochreiter and Schmidhuber [1997], the typical activation function units are replaced with "memory cells", where a pictorial representation of a cell is shown in Figure 7.2. The benefit of using LSTM units comes from the ability to store information, which means longer-range dependencies present in the data can be exploited. This is in comparison to standard RNNs which are only able to utilise short term information [Hochreiter and Schmidhuber, 1997]. Additionally, LSTM units are able to overcome the problems of vanishing gradients typically exhibited by standard recurrent neural networks. This behaviour is beneficial for speech processing applications as it allows for modelling of dynamically changing context present in a time-varying signal such as speech [Sak et al., 2014].

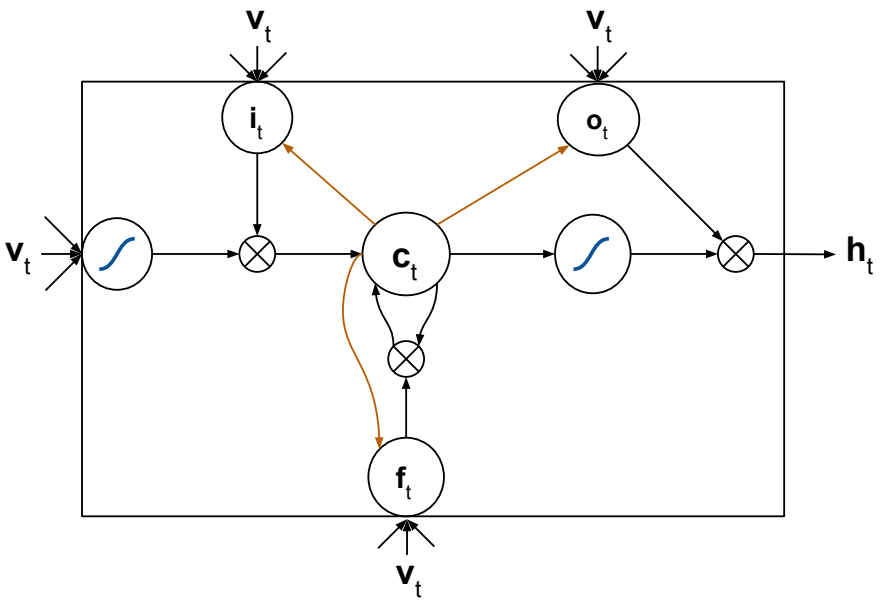



**Figure 7.2:** An LSTM cell showing the input gate, $\boldsymbol{i}_t$; output gate, $\boldsymbol{o}_t$; and forget gate, $\boldsymbol{f}_t$; which are used to control the centre storage cell, $\boldsymbol{c}_t$. Each input gate receives an input vector, $\mathbf{v}_t$, hidden layer outputs from the previous time-step, $\boldsymbol{h}_{t-1}$, and the storage cell value from the previous time-step, $\boldsymbol{c}_{t-1}$. The blue sigmoids indicate the tanh function. The orange lines show "peephole" connections which allow the gates to see what values are currently in the storage cell.

The LSTM units make use of gates to control the flow of input and output information from both the unit and the storage cell. There are three gates: input, output, and forget. The forget gate, $f_t$, makes decisions on what information currently in the storage cell should be kept or forgotten, and is described by,

$$\boldsymbol{f}_t = \sigma(\boldsymbol{v}_t \boldsymbol{W}_{vh} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{hf} + \boldsymbol{c}_{t-1} \boldsymbol{W}_{cf}), \tag{7.13}$$

where $v_t$ is the input data and $\boldsymbol{h}_{t-1}$ is the output from the hidden layer of the previous time-step. The $\boldsymbol{c}_{t-1}$ term is the information in the storage cell from the previous time-step, and provide "peephole" connections to the three gates. The subscript of the weight connections shows which data the connections are between, for example, $\boldsymbol{W}_{vh}$ are the input to hidden layer weights. When the output of the sigmoid activation, $\sigma$, is close to zero the information in the cell is forgotten, whereas when the output is close to one the previous information is retained.

Having decided what data should be forgotten, the new input data is determined and the storage cell is then subsequently updated. The input gate, $\boldsymbol{i}_t$, calculation is similar to that of the forget gate and is described by,

$$\boldsymbol{i}_t = \sigma(\boldsymbol{v}_t \boldsymbol{W}_{vh} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{hi} + \boldsymbol{c}_{t-1} \boldsymbol{W}_{ci}), \tag{7.14}$$

where outputs close to zero mean an input value will be ignored, and those close to one mean the inputs will be stored. The data in the storage cell, $\boldsymbol{c}_t$, can then be updated by applying,

$$\boldsymbol{c}_t = \boldsymbol{f}_t \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \tanh(\boldsymbol{v}_t \boldsymbol{W}_{vc} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{hc}), \tag{7.15}$$

where the forget gate outputs are used to remove information in the storage cell from the previous time-step, and the input gate outputs are used to indicate which

of the new values we should store in the cell.

Intuition behind the steps described thus far is as follows. The forget gate controls what the storage cell currently "knows", and if no new information is presented then the contents of the cell are not going to be updated. However, given some new information in the input data, the forget gate controls the removing of the old information allowing for the new information to be stored in the cell.

Finally, the output gate, $\mathbf{o_t}$, is used to decide what information should be output from the storage cell, $c_t$. Values from the output gate are given by,

$$\boldsymbol{o}_t = \sigma(\boldsymbol{v}_t \boldsymbol{W}_{vh} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{ho} + \boldsymbol{c}_t \boldsymbol{W}_{co}), \tag{7.16}$$

where the primary difference between the output gate equation, and those of the input and forget gates, is that the output gate is only provided with the current storage cell information, and the information from the previous time-step. To get the output from the LSTM unit, $\boldsymbol{h}_t$, the following is calculated,

$$\boldsymbol{h}_t = \boldsymbol{o}_t \cdot \tanh(c_t), \tag{7.17}$$

where the final application of the tanh function ensures the outputs from the unit are in the range of $-1$ to 1.

### 7.3.2 Bi-directional layers

In typical RNN architectures, only past information is used to decide upon the current network output. However, it has been found that by also including future information the performance can be further improved over uni-directional models [Graves et al., 2013a]. Bi-directional recurrent layers can be formed by using two hidden layers where one computes the forward hidden sequence, $\overrightarrow{\mathbf{h}}$, and

the other computes the reverse hidden sequence, $\overleftarrow{\mathbf{h}}$. Element, $y_t$, of the output sequence can be then be obtained through application of

$$y_t = \mathbf{W}_{\overrightarrow{h}y}\overrightarrow{\mathbf{h}}_t + \mathbf{W}_{\overleftarrow{h}y}\overleftarrow{\mathbf{h}}_t. \tag{7.18}$$

### 7.3.3 Network architecture and training

To exploit the ability of deep neural network architectures to extract higher-level representations of the input data, multiple bi-directional LSTM layers can be stacked together. For implementation purposes, this means the addition of another hidden layer for each forward hidden layer, with processing performed backwards in time, such that the processing would begin at element $T$ in the sequence, and work backwards to element one.

Training of recurrent neural networks is performed using the backpropagation through time technique, which is based on the standard backpropagation of errors method used to train feed-forward neural networks. As with standard neural networks, optimisation techniques such as stochastic gradient descent and resilient backpropagation can be used. Aside from issues of vanishing gradients, it is possible that computed gradients may "explode," whereby the values become extreme large. A simple solution to this issue is to clip the gradients at a pre-defined threshold [Pascanu et al., 2013].

The deep bi-directional LSTM (DB-LSTM) architecture used in this work follows that of Graves et al. [2013a], where each of the bi-directional layers consists of 500 LSTM units (250 units for the forward layer, and 250 for the backward layer), and three bi-directional layers are stacked together between the input and output layers. A batch size of 512 examples is used, with a gradient clipping value of one. Gaussian noise is added to the weight parameters as a form of regularisation,

and training is completed once validation scores converge with no further increase in classification accuracy observed. More details of the architecture are given in Appendix B.5.

The application of DB-LSTMs for performing visual-to-audio domain mapping is explored in the next section. First, experiments are conducted to determine an optimal audio and visual sequence length, $T$, and then objective intelligibility investigations are performed to see which of the two model-level methods perform best.

## 7.4 Evaluation

In this section, the performance of the two model-level visual-to-audio mapping methods, incorporating temporal information, is explored. First, an investigation is conducted for the DB-LSTM system to determine an optimum sequence length, $T$, of input visual vectors and corresponding output codebook entry labels. Then, experiments are conducted for the Viterbi decoding method by applying different weightings to the emission and transition probabilities to determine an optimal combination, by exploring different values of $\gamma$ in Equation 7.10. The mean squared error between the original and estimated Mel-filterbank features is recorded for both of these experiments, with a subset of the configurations from each method selected for further testing. Reconstructed utterances, using the monotone artificial-$f_0$ contour with the joint aperiodicity codebook estimates, are evaluated objectively using the STOI and PESQ measures. The best performing method is then investigated further with subjective listening tests, discussed in Chapter 8. Finally, an analysis is presented to understand the characteristics of reconstructed audio speech signals.

### 7.4.1 LSTM sequence length

To find an optimum sequence length for estimating the audio codebook entry labels from input visual feature vectors, varying sequence lengths are investigated in the DB-LSTM models. Sequences of $T = \{3, 7, \ldots, 31, 35\}$ are explored to see which length gives the minimum mean squared error between the estimated and original Mel-filterbank audio features. The visual sequences are comprised of contiguous AAM visual features, which are used to predict sequences of contiguous codebook entry labels. During the training phase of the DB-LSTM, the network will learn the temporal relationship between corresponding sequences of codebook entry labels and visual feature vectors. To produce the final output audio feature estimates, the third method discussed in Section 6.4.2 is applied to perform an overlap-and-add of the audio features pertaining to the individual sequences.

**Table 7.1:** Mean squared error between audio feature estimates from the DB-LSTM and the original Mel-filterbank features, for the female speaker, with varying sequences lengths, $T$.

| Sequence length, T | MSE |
|---|---|
| 1 | 0.938 |
| 3 | 0.728 |
| 7 | 0.560 |
| 11 | 0.481 |
| 15 | 0.442 |
| 19 | 0.418 |
| 23 | 0.404 |
| 27 | 0.399 |
| 31 | 0.390 |
| 35 | 0.381 |

Mean squared errors are reported in Table 7.1 for the various sequence lengths. As the length of the sequences increases, the audio feature vector estimates im-

prove, where the DB-LSTM model with a sequence length of $T = 35$ (covering 360 ms of audio and visual signal) gives the lowest error. A subset of these configurations, with sequence lengths of $T = \{23, 27, 31, 35\}$ are further evaluated using objective intelligibility measures to identify the best performing system.

### 7.4.2 Viterbi matrix weightings

In the standard application of the Viterbi algorithm, the transition and emission probabilities are given equal weighting. However, the contribution of the probabilities of each matrix can be weighted, so as to ascribe more importance to one over the other. Variable $\gamma$, in Equation 7.10, performs this weighting function, where $\gamma = 0$ means only the emission probabilities will be used, and $\gamma = 1$ means only the transition probabilities will be used. With $\gamma = 0.5$ both the emission and transition probabilities are given equal weight. Accordingly, experiments are performed comparing the MSE between original audio features and those reconstructed using the Viterbi algorithm with different matrix weightings applied.

In Table 7.2 it can be seen that as the transition probabilities are given more weight there is a slight decrease in MSE up to $\gamma = 0.3$, after which the error increases quickly up to $\gamma = 1.0$. The lowest MSE is achieved when using $\gamma = 0.3$ with an error of 0.919. In comparison to Table 7.1, the MSEs for the Viterbi method are all in the region of the error achieved for the DB-LSTM system when using a sequence length $T = 1$, which is equivalent to a single frame system.

The transition matrix produced for an audio codebook of size 1024 has over 1 million values ($1024 \times 1024$). This matrix is populated from approximately 240,000 codebook entry pairs (300 frames per utterance and for 800 training utterances), which causes the matrix to be extremely sparse and likely to be dominated by a small number of codebook entry pairs. For the transition matrix to be accurate, a

**Table 7.2:** Mean squared error of original and estimated audio features from the female speaker using the Viterbi method with different weightings, $\gamma$, for the transition matrix, $\boldsymbol{A}$, and emission matrix, $\boldsymbol{B}$.

| Gamma, $\gamma$ | MSE |
|---|---|
| 0.0 (Emission only) | 0.927 |
| 0.1 | 0.923 |
| 0.2 | 0.922 |
| 0.3 | 0.919 |
| 0.4 | 0.928 |
| 0.5 (Equal weight) | 0.941 |
| 0.6 | 0.966 |
| 0.7 | 1.024 |
| 0.8 | 1.166 |
| 0.9 | 1.590 |
| 1.0 (Transition only) | 3.114 |

considerable amount more data are required to generate it, and the amount of data used is simply not adequate to give good accuracy in the estimated audio features. Accordingly, poor MSE scores between the original and estimated Mel-filterbank features are observed.

The application of Viterbi decoding to the estimated probabilities of the codebook entry labels using first-order transitions, in an attempt to incorporate longer-range temporal information, does not show any real benefit. Additionally, using the class priors from the training data in Equation 7.1 resulted in lower errors than when using uniform probabilities for $P(c_j)$, in effect meaning that $P(\mathbf{v}|c_j) = P(c_j|\mathbf{v})$. Objective evaluations are conducted on the configurations with weighting values of $\gamma = \{0.1, 0.2, 0.3\}$.

### 7.4.3 Objective experiments

Objective intelligibility scores, using STOI and PESQ, of reconstructed audio utterances are now evaluated for both methods proposed in this chapter. In Section 7.2, the Viterbi algorithm is used to produce a sequence of codebook entry labels using estimated probabilities from a neural network and first-order class label transitions. The optimum weighting, $\gamma$, for the probability matrices which resulted in the lowest MSE, was found to be in the range of 0.1–0.3, with the best performance at $\gamma = 0.3$. The three configurations of the Viterbi method that gave the lowest error are evaluated in this section for the female speaker. For the DB-LSTM system, presented in Section 7.3, it was found that the MSE decreased as the sequence length increased, i.e. as longer-range temporal information was incorporated into the model. Four sequence lengths of $T = \{23, 27, 31, 35\}$ for input visual features and output codebook entry labels are evaluated here using objective measures for the male and female speakers.

**Table 7.3:** STOI and PESQ results for utterances reconstructed using the Viterbi method with three values of $\gamma$ for the female speaker.

| $\gamma$ | STOI | PESQ |
|---|---|---|
| 0.1 | 0.456 | 0.923 |
| 0.2 | 0.464 | 0.923 |
| 0.3 | **0.470** | **0.925** |

Table 7.3 shows the STOI and PESQ scores for utterances reconstructed for the female speaker using the Viterbi decoding method. Confirming the MSE results, the best performing system was found when using a weighting of $\gamma = 0.3$. That is, the transition probabilities have a contribution of 0.3, and the emission probabilities have a contribution of 0.7. However, both the STOI and PESQ scores are low, indicating that the method leads to reconstructed utterances with poor intelligibility. In comparison, the best performing clustering-and-classification method,

using feature-level temporal encoding, from the previous chapter (see Section 6.5), achieved a STOI of 0.740, and PESQ of 1.668, for the female speaker.

Table 7.3 shows the STOI scores for reconstructed utterances from the DB-LSTM for the male and female speakers using sequence lengths of $T = \{23, 27, 31, 35\}$. It can be seen that the best scores are obtained when using a sequence length of $T = 35$, with 0.704 for the female speaker and 0.694 for the male. Confirming the MSE analysis from Table 7.1, the performance of the DB-LSTM improves when using longer sequence lengths. Furthermore, the female exhibits slightly higher predicted intelligibility over the male speaker.

**Table 7.4:** STOI intelligibility scores for utterances reconstructed from the DB-LSTM method for the female and male speakers.

| $T$ | Female | Male |
|---|---|---|
| 23 | 0.688 | 0.680 |
| 27 | 0.693 | 0.687 |
| 31 | 0.701 | 0.687 |
| 35 | **0.704** | **0.694** |

Table 7.5 shows PESQ scores for utterances reconstructed using the DB-LSTM estimations for both the female and male speakers with various sequence lengths. As with the STOI results, the highest PESQ scores are achieved with a sequence length of $T = 35$, with a score of 1.550 for the female speaker and 1.830 for the male. In comparison to the STOI results, it is found that the male exhibits higher PESQ scores over the female.

From these objective intelligibility experiments it is evident that the DB-LSTM performs considerably better than the Viterbi decoding method. For the female speaker, the best performing Viterbi method achieved a STOI of 0.470 and PESQ of 0.925, where both scores are much lower than those obtained for the DB-LSTM method, with 0.704 and 1.550 for STOI and PESQ, respectively. In comparison to

**Table 7.5:** PESQ scores for utterances reconstructed using the DB-LSTM method for the female and male speakers.

| $T$ | Female | Male |
|---|---|---|
| 23 | 1.517 | 1.759 |
| 27 | 1.527 | 1.823 |
| 31 | 1.542 | 1.822 |
| 35 | **1.550** | **1.830** |

the best performing method from the previous chapter, the STOI and PESQ results are slightly lower for both speakers, suggesting that encoding temporal information at the model-level is less favourable than at the feature-level. However, despite this, the results still show the benefits of incorporating longer-range temporal information (above 300 ms) when producing audio estimates from visual speech.

For both the male and female speakers, the highest objective intelligibility results were obtained when using a sequence length of $T = 35$. However, as the errors are similar, a sequence length of $T = 31$ is chosen for the DB-LSTM to remain in keeping with the window sizes chosen for the feature-level configuration discussed in Section 6.4. This system is investigated further in Chapter 8, where subjective intelligibility experiments are conducted on reconstructed utterances from the female speaker.

### 7.4.4 Utterance analysis

To evaluate further the performance of the proposed Viterbi decoding and DB-LSTM methods, an audio analysis is performed on reconstructed utterances from both methods.

Wideband spectrograms are shown in Figure 7.3 for an original utterance and of reconstructed utterances using audio estimates from the two model-level methods.

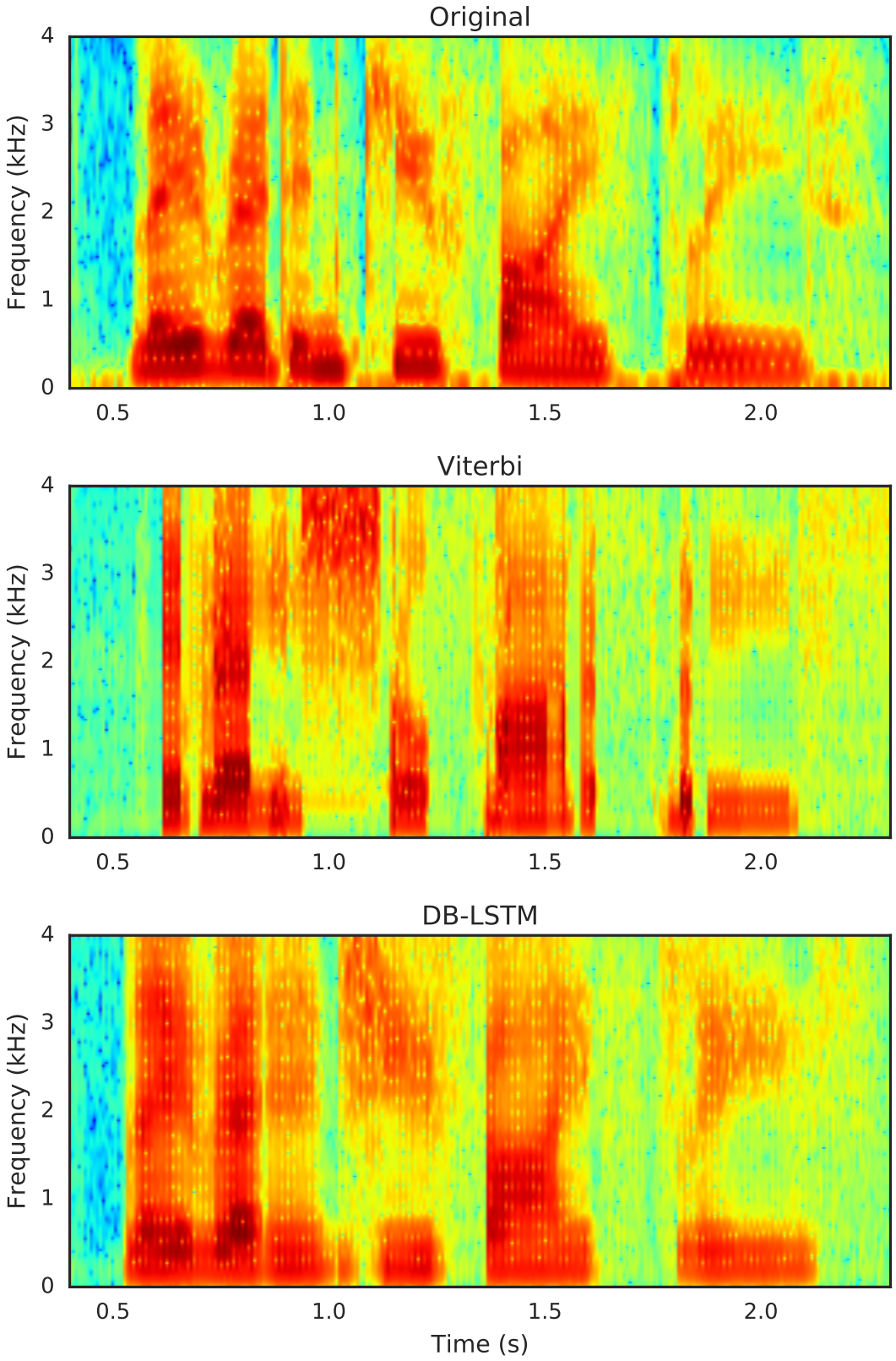



**Figure 7.3:** Wideband spectrograms of the original utterance "lay red in Q 5 please" and reproductions from the Viterbi method ($\gamma = 0.3$) and DB-LSTM ($T = 31$) visual-to-audio domain mapping models for the female speaker.

The spectrogram for the DB-LSTM appears relatively faithful to the original with noticeable high-frequency spectral detail, where the formants $F_2$ and $F_3$ can be seen in addition to the first formant, especially between 0.5–1 s. However, smoothing of the spectral-envelope can be observed. The spectrogram of the utterance reconstructed from the Viterbi decoding method shows a number of artefacts in comparison to the original, where there are blocks of spectral-envelope regions with obvious beginning and end points that do not occur in the original signal. From informal listening tests, these artefacts are especially apparent where erratic changes between these regions of the spectral-envelope adversely affect the reconstructed audio signals.

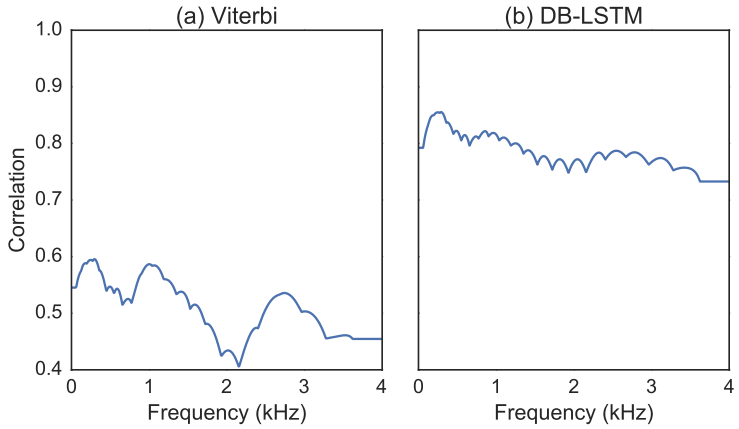



**Figure 7.4:** Correlations of frequency bins between original and estimated spectral-envelope surfaces for the Viterbi and DB-LSTM models, respectively.

Figure 7.4 shows correlations over frequency between original and estimated spectral-envelope surfaces for the test utterances of the female speaker for both model-level methods. It is readily apparent that the DB-LSTM exhibits considerably higher and more uniform correlations across all frequencies over the Viterbi method. The frequency correlations for the DB-LSTM system are close to those shown in Figure 6.8 for the feature-level clustering-and-classification approach, where the slightly lower correlations are expected based on the lower STOI and

PESQ scores observed. In comparison, the frequency correlations for the Viterbi method are far lower and more variable, ranging from 0.4–0.6, which confirms the objective intelligibly results and the observations of the spectrogram suggesting that the method yields poor audio feature estimates.

## 7.5 Summary

In this chapter, two methods are explored for incorporating temporal information at the model-level. In the first method, Viterbi decoding is applied to codebook entry probabilities estimated from visual features using a DNN. First-order transitions are determined from the training data and used in the algorithm, with the output being the sequence of codebook entry labels that has the highest cumulative probability. Furthermore, weighting can be applied to the emission and transition probabilities to ascribe more importance to one over the other, and vice-versa. The second method uses a recurrent neural network with the long short-term memory architecture to estimate sequences of codebook entry labels from corresponding sequences of input visual features. Stacking of multiple bi-directional layers is performed to give a deeper neural network architecture. To obtain an output spectral-envelope time-frequency surface, the overlap-and-add method (see Section 6.4.2) is applied to the Mel-filterbank features pertaining to the sequences estimated by the DB-LSTM.

Experiments are first conducted to determine an optimum sequence length, $T$, for the DB-LSTM, and an optimum weighting, $\gamma$, for the Viterbi method. Objective intelligibility evaluations show that, while the scores are not as high as when using the feature-level clustering-and-classification methods detailed in Chapter 6, incorporating temporal information at the model-level yields good results for the DB-LSTM when using a sequence length of $T = 35$. To remain in keeping with the

configurations selected for the best performing feature-level method, which used audio and visual windows of $S^A = S^V = 31$, the DB-LSTM configuration is chosen such that $T = 31$. In comparison, although the Viterbi decoding method exhibits a reduction in MSE as the weight of the transition probabilities is increased, up to $\gamma = 0.3$, the method still performs poorly, resulting in very low STOI and PESQ scores.

In the next chapter, intelligibility results from subjective listening tests of utterances reconstructed using the best-performing regression, feature-level, and model-level methods are evaluated. Additionally, experiments are presented on a larger-vocabulary dataset to determine the performance of the best system when applied to more continuous and less constrained speech.

# Chapter 8

# Evaluation

## 8.1 Introduction

This chapter is concerned with the subjective evaluation of speech intelligibility for utterances reconstructed using audio estimates from the regression, feature-level, and model-level approaches. Subjective intelligibility tests are conducted for the female speaker from the GRID dataset using the best configurations from each of the three approaches, using audio-only and audiovisual media. Investigations are then conducted for speech reconstructions of utterances from a dataset with a larger-vocabulary and less-constrained speech.

Results from the subjective listening tests for the regression system were presented in Chapter 5, showing that word-level accuracies, the number of correctly identified words within an utterance, were significantly higher than chance accuracy at 19 %. However, the intelligibility of the best performing system, using a GMM to estimate LPC coefficients from AAM features, was similar to that of the visual-only utterances (where untrained listeners are asked to perform lip-reading) at around 50 %. Two issues identified with the regression approach are

that non-plausible spectral-envelopes are generated from the estimated real-valued and continuous audio features, and that the audio estimates are produced from only a single visual vector, with no temporal information. Accordingly, to incorporate temporal information, a clustering and classification framework with feature-level temporal encoding is presented in Chapter 6, and with temporal encodings at the model-level in Chapter 7. Both proposed systems result in considerably lower mean squared error between the original and estimated Mel-filterbank features, and much improved scores for the STOI and PESQ objective measures. This suggests that the intelligibility of the reconstructed audio utterances is far higher than the original regression system. In this chapter, the feature-level and model-level systems, in addition to the regression system, are evaluated using subjective listening tests with human listeners to determine the intelligibility of speech utterances reconstructed using audio estimates from the three approaches.

The GRID corpus was chosen as the main dataset for this work as the highly-constrained grammar and small vocabulary size makes the transcription task easier for listeners and allows for usable intelligibility scores to be obtained. Developing the methods for a larger dataset would have been problematic as it would likely be the case that word-level accuracies would be so low that it would be very difficult to discern between the relative performance of different systems. The results of the work presented in this thesis thus far have shown promise for being able to reconstruct intelligible audio speech from visual information. Accordingly, the best performing method identified from the subjective listening tests in the next section is applied to reconstruct utterances for a male speaker from a dataset with a larger-vocabulary and less-constrained speech. Objective evaluations are conducted, in addition to an audio analysis, to see how well the proposed methods perform on this bigger dataset.

The remainder of this chapter is organised as follows. In Section 8.2, subjective

listening test results for the feature-level and model-level approaches, including the regression system as a baseline, are presented for the female speaker from GRID, with additional analyses of reconstructed utterances. Experiments on a larger-vocabulary dataset are investigated in Section 8.3, to see how the best performing model performs on larger, less-constrained audiovisual speech corpora. Lastly, the results in this chapter are summarised in Section 8.3.

## 8.2 GRID subjective evaluations

In this section, the results of subjective intelligibility tests are presented for the utterances reconstructed for the female speaker from GRID using the three main visual-to-audio mapping methods. Mean squared error and objective tests have been used to identify configurations that perform well for the approaches presented in Chapter 6 and Chapter 7. These best performing configurations are now analysed subjectively through human listening tests to determine word-level accuracies. Three systems are evaluated:

1. REG: the baseline regression system from Chapter 5, where a DNN with a linear output layer is used to estimate single frames of real-valued Mel-filterbank audio features from single frames of input AAM visual feature vectors (2D-DCT visual features were used in the original system),

2. FLE: the clustering and classification system using feature-level windows with size $S^A = S^V = 31$ for incorporating longer-range temporal information, as presented in Chapter 6,

3. MLE: a recurrent neural network system using the long short-term memory architecture from Chapter 7, with a sequence length of $T = 31$ for incorporating temporal information at the model-level.

The window sizes, $S^A$ and $S^V$, and sequence length, $T$, were all selected to be 31 vectors in length for consistency, and resulted in configurations that give close to the best objective performance across all systems. To investigate the usefulness of adding the original visual stream, two further configurations are explored to measure the intelligibility of the feature-level and model-level systems with the listeners also able to watch the video in addition to hearing the reconstructed audio speech. These audiovisual test configurations are referred to as FLE+V and MLE+V for the feature-level and model-level systems, respectively. Audiovisual evaluations for the regression system were presented previously in Chapter 5, and, as such, are not performed again here. Finally, a further test (VIS) is included to measure the intelligibility when subjects were presented with just the original video stream, where listeners are required to perform lip-reading.

The listening tests were performed with twenty listeners, who were located in a quiet room and used headphones to listen to the reconstructed utterances. Speech from the female speaker from the GRID dataset was used, with 750 sentences for training and the remaining 250 sentences used for testing. The monotone method provided artificial-$f_0$ contours, and with aperiodicity information estimated using the joint-feature clustering approach. The listening tests were conducted using the web-based interface, as discussed in Section 5.2 and shown in Figure 5.6. Each listener was presented with four utterances from each of the six test configurations (REG, FLE, MLE, FLE+V, MLE+V, and VIS), hearing 24 utterances in total, and was allowed to listen to each utterance as many times as they desired. The question order was randomised and the sentences selected such that each listener would only hear one occurrence of an utterance. The listeners used the drop-down boxes to select their choices for each word, and intelligibility is calculated on a per-word basis. For all of the listeners, the four repetitions for each configuration were grouped to give one accuracy score per configuration.

### 8.2.1 Listening test results

Table 8.1 shows the intelligibility (word-level accuracy) scores averaged across all twenty subjects for the six test configurations. The results show that the proposed clustering-and-classification approaches (FLE and MLE), incorporating longer-range temporal information, produce speech of substantially higher intelligibility than the baseline regression method (REG). As was found with both the MSE analysis and objective tests, the subjective tests show that the feature-level estimation produces more intelligible speech than the model-level estimation: 77.1 % in comparison to 74.4 %. Including the original video signal in the tests (FLE+V and MLE+V) results in a further increase in intelligibility of around 7 % in both cases. The intelligibility of the visual-only signal (lip-reading) was substantially lower at 47.5 %, although better than the audio-only regression system (REG). Additionally, all of the evaluated systems achieve intelligibilities higher than chance, which is 19 % for the GRID grammar (see Equation 5.12).

**Table 8.1:** Word accuracy scores (and standard error) from subjective listening tests showing the intelligibility of each configuration.

| Configuration | Name | Accuracy (%) |
|---|---|---|
| Regression | REG | 30.4 (2.9) |
| Feature-level | FLE | 77.1 (2.6) |
| Model-level | MLE | 74.4 (1.9) |
| Feature-level with video | FLE+V | 84.2 (1.8) |
| Model-level with video | MLE+V | 80.8 (1.7) |
| Video-only | VIS | 47.5 (3.3) |

The highest overall intelligibility of 84 % was achieved for utterances reconstructed using the FLE+V configuration, where the reconstructed utterances were combined with the original video signal. That is, the feature-level clustering and classification method was used for estimating the spectral-envelope parameter,

aperiodicity information was obtained from the joint-feature estimation method, the monotone artificial-$f_0$ method was used to provide the fundamental frequency contour, and STRAIGHT was used for reconstructing the audio speech.

### 8.2.2 Analysis of confusions

To further understand the ability of the models to yield intelligible speech reconstructions, a more in depth analysis is conducted on the accuracies achieved for each category of the GRID grammar. From Table A.1 it can be seen that there are a greater numbers of choices for the letters (25) and digits (10), compared to the other categories which each have four choices (command, colour, preposition and adverb). To investigate the effect of this further, Table 8.2 shows word-level accuracy for each grammar category for each of the six configurations tested.

**Table 8.2:** Per-word accuracy scores for each of the six different system configurations.

| | REG | FLE | MLE | FLE+V | MLE+V | VIS |
|---|---|---|---|---|---|---|
| Command | 22.5 | 92.5 | 91.2 | 98.8 | 93.8 | 52.5 |
| Colour | 50.0 | 96.2 | 98.8 | 96.2 | 98.8 | 66.2 |
| Preposition | 26.2 | 72.5 | 58.8 | 77.5 | 76.2 | 42.5 |
| Letter | 8.8 | 16.2 | 21.2 | 36.2 | 27.5 | 8.8 |
| Digit | 27.5 | 88.8 | 85.0 | 96.2 | 90.0 | 43.8 |
| Adverb | 47.5 | 96.2 | 91.2 | 100.0 | 98.8 | 71.2 |

The table reveals significant variation in word-level accuracy between the categories. Considering the best performing feature-level method (FLE), word accuracy for command, colour and adverb categories is over 90 %, while for prepositions it is lower at 73 %. Digit accuracy is also high at nearly 36 %, while letter accuracy is considerably lower although still higher than that of chance. A similar trend is observed for the MLE system, and overall higher accuracies are found

for the audiovisual media systems. Interestingly, the adverb and colour categories for the visual-only utterances (VIS) show intelligibility scores of approximately 70 %, which appears to be very high given that the listeners were all untrained lip-readers. Whilst this does suggest that the GRID dataset is somewhat easy and not a very realistic example, it is important to remember that the dataset was chosen as the task of reconstructing audio from visual speech is difficult.

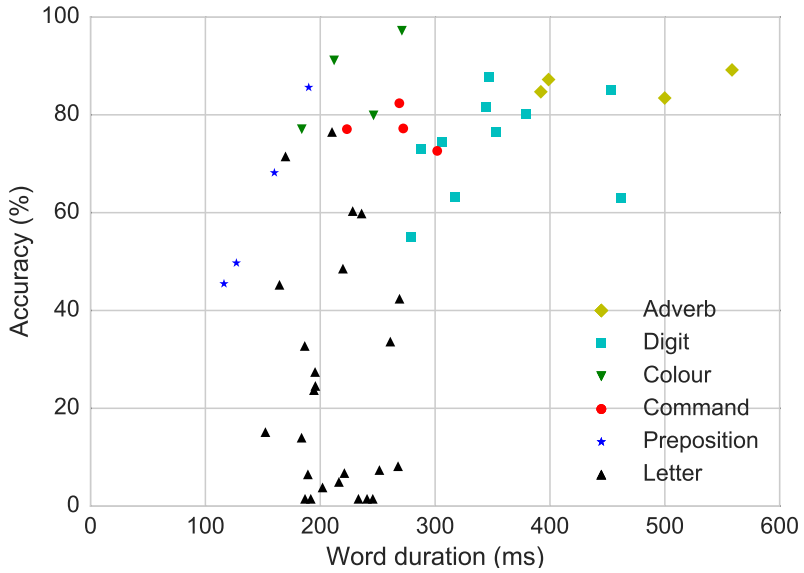



**Figure 8.1:** Scatter plot of average word duration and word accuracy, broken down into GRID grammar categories.

The number of choices within each category has an effect on word accuracy, yet it also speculated that word duration plays a significant factor. Accordingly, Figure 8.1 shows a scatter plot of the mean duration of each word in the GRID grammar (averaged over all occurrences of that word) against the mean word accuracy for that word in the subjective listening tests. Considering first the word accuracy of the preposition category, which is lower than that of the command, colour and adverb categories, these words have durations less than 200 ms which is considerably shorter than the command, colour and adverb words, which are,

in general, longer and have higher accuracy. The letters category, which show the lowest accuracy, are all of a short duration of roughly 200 ms. In fact, words with a duration over 300 ms are all recognised with greater than 60 % accuracy.

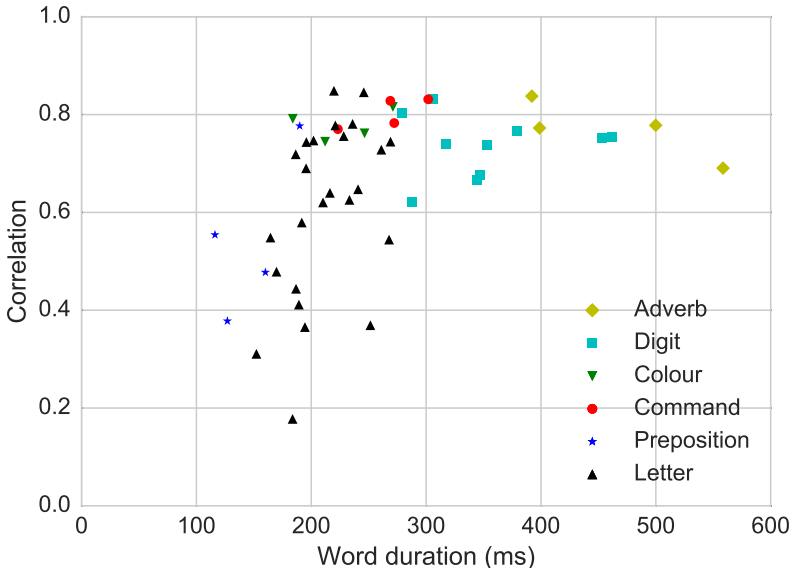



**Figure 8.2:** Scatter plot of average word duration and average correlation between the original and estimated spectral-envelopes for each word, broken down into GRID grammar categories.

To further understand the observations between average word duration and accuracy, the average spectral-envelope correlations between the original and estimated surfaces for each word of the tests utterances are plotted against word duration in Figure 8.2. This investigation is performed to explore whether there exists a relationship between word duration and how precisely the spectral-envelope for the words have been estimated. The correlations show a similar trend to the accuracies observed in Figure 8.1, where it can be seen that words with a duration over 300 ms all have correlations greater than 0.65, and consists of words from the adverb and digit categories. The majority of words with a duration between 200–300 ms show correlations greater than 0.6, and those below 200 ms show correlations less than 0.6. Interestingly, the letters show stronger correlations than

would be expected given the accuracies with which they are recorded.

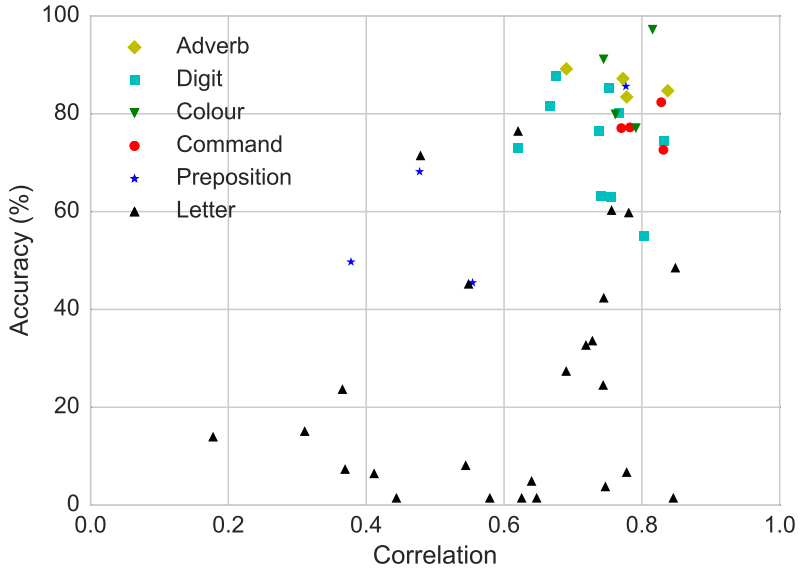



**Figure 8.3:** Scatter plot of word accuracy and average correlation between the original and estimated spectral-envelopes for each word, broken down into GRID grammar categories.

To further explore this, the average word correlations are shown against accuracy in Figure 8.3. In this figure it can be seen that the majority of the letters show accuracies lower than 40 %, yet the correlations range from 0.15–0.9, indicating that there are a considerable number of confusions for the letters category. The words in the adverb, digit, command, and colour categories are identified with greater than 60 % accuracy, showing correlations of higher than 0.6. The results from this experiment suggest that whilst the average spectral-envelope correlation for some words may be high, the correlation alone does not necessarily provide a strong indication as to whether or not the word may be intelligible. However, it appears that stronger correlations are suggestive of better intelligibility for words from categories with fewer choices, and for those with longer durations.

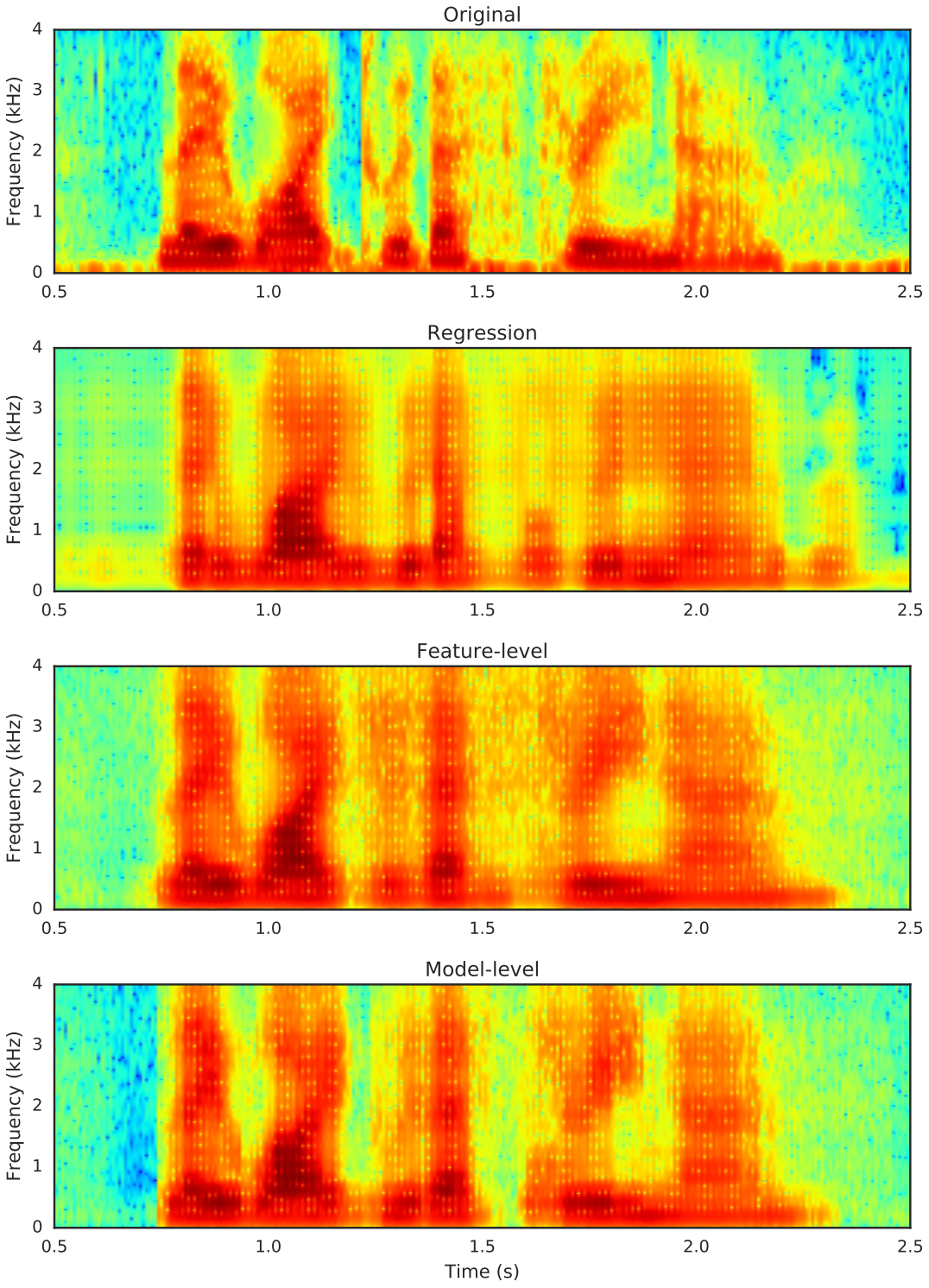



**Figure 8.4:** Wideband spectrograms of the sentence "*lay white with F 3 now*" spoken by the female speaker, for the original utterance, and reconstructed utterances using the regression, feature-level, and model-level systems.

### 8.2.3 Spectrogram analysis

To illustrate the audio information extracted from the visual speech features, Figure 8.4 shows wideband spectrograms of the sentence "*lay white with F 3 now*" taken from the female speaker, for the original signal and for those reconstructed using the regression, feature-level and model-level methods (REG, FLE, and MLE). For the three reconstructed utterances, the overall energy and voice activity of the speech signals closely matches that of the original. Furthermore, confirming the objective and subjective test results, the feature-level and model-level spectrograms appear more similar to those of the original and show a significantly better representation of the formant structure than the regression method. However, one clear artefact of the reconstructed speech is the widening of formant bandwidths compared to the original speech.

## 8.3 Larger dataset

In the previous section, the best performing visual-to-audio system for the GRID corpus was the feature-level method (FLE), achieving an audio-only intelligibility of 77 %, and an audiovisual accuracy of 84 %. In this section, the application of the feature-level method is explored for the larger-vocabulary and less-constrained RM-3000 dataset. This investigation is conducted to determine how well the visual-to-audio approaches, and methods for synthesising excitation, generalise to bigger audiovisual speech datasets for the reconstruction of intelligible audio speech.

The RM-3000 audiovisual dataset collected by Howell [2015], contains 3000 sentences selected from the Resource Management corpus spoken by a native English speaker, and with a vocabulary size of around 1000 words. Pre-extracted AAM features are provided and used as the visual input to estimate windows of

Mel-filterbank audio features, as has been described previously in Chapter 6 for the clustering and classification system with feature-level temporal encoding. In total, there is 260 minutes of data, with 2250 utterances used for training, and the remaining 750 used for testing. Further details about the dataset are provided in Appendix A.2. For reference, utterances are reconstructed using the monotone artificial-$f_0$ method with a mean value of $f_{0_i} = 185$, and with aperiodicity information estimated using the joint-feature clustering approach.

**Table 8.3:** Mean squared error between the original and estimated Mel-filterbank amplitudes with varying audio and visual window sizes.

| $S_A$ \ $S_V$ | 7 | 15 | 23 | 31 |
|---|---|---|---|---|
| 7 | 0.694 | 0.639 | 0.625 | 0.619 |
| 15 | 0.586 | 0.545 | 0.539 | **0.530** |
| 23 | 0.697 | 0.662 | 0.624 | 0.609 |
| 31 | 0.678 | 0.639 | 0.622 | 0.613 |

Various audio and visual window sizes, $S^A = S^V = \{7, 15, 23, 31\}$, are explored to find an optimum combination of the two that gives the lowest MSE between the original and estimated Mel-filterbank features. Table 8.3 shows the MSEs for the various audio and visual window sizes. It can be seen that, for each of the audio window sizes, there is a reduction in MSE as the size of the visual window increases from 7 to 31. However, the same reduction in error is not observed as the size of the audio windows increases. Interestingly, the audio window size that gives the lowest error for all visual windows is $S^A = 15$, with the lowest overall MSE achieved when using a visual window of $S^V = 31$. This configuration is further explored using objective measures of speech intelligibility.

Utterances are reconstructed using the feature-level method of Mel-filterbank estimation with window sizes of $S^A = 15$ and $S^V = 31$ for the audio and visual features, respectively. Objective intelligibility measures applied to the set of re-

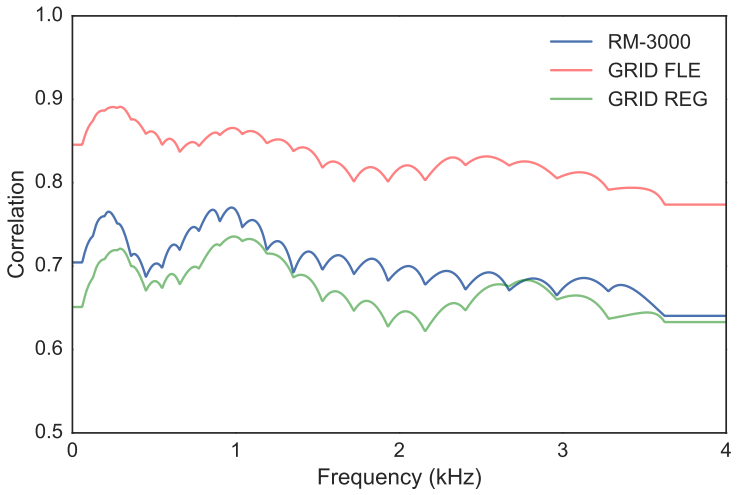



**Figure 8.5:** Correlations of frequency bins between the original and estimated spectral-envelope surfaces for the feature-level system applied to the RM-3000 dataset, and for comparison: the FLE and REG systems for GRID.

constructed test utterances show scores of 0.612 for STOI, and 1.693 for PESQ. These values are lower than those obtained for the feature-level approach for the male speaker from the GRID dataset, at 0.735 and 2.055 for STOI and PESQ, respectively. Additionally, the scores are similar to those obtained for the regression system for the male speaker where a STOI of 0.604 and PESQ of 1.700 was recorded. Informal listening tests suggest that, whilst there is a noticeable speech-like quality to the utterances, and broad spectral detail seems to be evident, the intelligibility of the utterances is low. Accordingly, subjective listening tests were not performed as it is likely that little usable information would be obtained from the experiments.

To further explore the lower intelligibility of the utterances, Figure 8.5 shows the correlation across frequency for all test utterances between the original spectral-envelopes and those reconstructed from the estimate Mel-filterbank features. For comparison, the correlations of the FLE and REG approaches for GRID, from Fig-

ure 6.8, are also shown. It can be seen that the correlations for RM-3000 are lower than those exhibited for the feature-level method as applied to the GRID dataset, yet the correlations are slightly higher than those shown for the regression system. There are prominent peaks at 250 Hz and 1 kHz, and the correlation then decreases beyond 1.4 kHz. Interestingly, there does not seem to be a peak in the 2.5–3 kHz region as is observed in the correlation analysis for the regression and feature-level methods, suggesting that the utterances lack sufficient high-frequency detail. Furthermore, this observation, and the uniformly lower correlations across frequency, provides evidence as to why the RM-3000 utterances are not as intelligible as those from the GRID corpus.

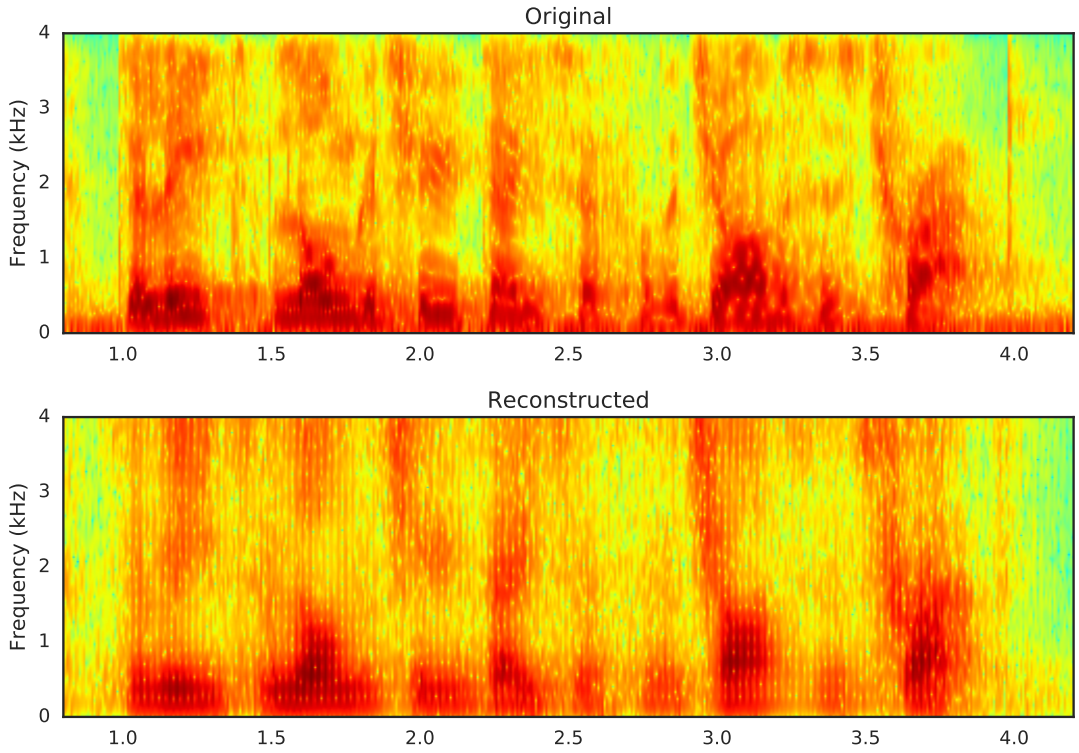



**Figure 8.6:** Wideband spectrograms of the utterance "*delete longitude data for the Jarvis's track*" showing original and reconstructed utterances.

The reduction in overall and high-frequency spectral-detail can be seen in the

spectrograms for an original and reconstructed utterance shown in Figure 8.6. The effect is similar to that which can be seen in Figure 8.4 for utterances reconstructed using the regression approach on the GRID dataset. In the reconstructed utterance, there is little evidence of formant structure other than the first formant, $F_1$, where the higher-frequencies appear to lack considerable detail. Some formant structure is noticeable around 2.5 s and 3.4 s, although it is apparent that the formant bandwidths are somewhat broader than those shown for the original utterance.

Whilst the intelligibility for RM-3000 is lower than that observed for GRID, the correlation analysis indicates that the visual-to-audio mapping still shows a benefit for speech processing applications on larger datasets where audio features are estimated from visual speech.

## 8.4 Summary

The subjective intelligibility experiments presented for the female speaker from the GRID dataset in this chapter, show that intelligible audio speech signals can be reconstructed from visual speech information using the feature-level and model-level approaches to visual-to-audio mapping. Specifically, the proposed feature-level clustering-and-classification method using deep neural networks achieves an intelligibility of 77 %, which is significantly better than the baseline regression method presented in Chapter 5 which had an intelligibility of 30 %. Additionally, the model-level DB-LSTM method achieved accuracies of 74 %. Incorporating longer-range temporal information was found to be important in producing intelligible speech reconstructions, with the best performance achieved using audio and visual window widths of around 300 ms in duration. Supplementing the audio signal with the video information in the subjective intelligibility tests gave further im-

provement, increasing the intelligibility by around 7 % This difference between the audio-only and audiovisual media suggests that there is still more information that can be extracted from the visual features. Furthermore, it was discovered that words with a longer duration tended to be recognised with greater accuracy than those with a shorter duration.

The experiments conducted on the larger RM-3000 dataset showed lower objective intelligibility scores than those obtained for the GRID dataset. Reconstructed utterances yielded STOI scores of 0.612 and scores of 1.693 for PESQ. Informal listening tests indicate the reconstructed utterances have a speech-like characteristic, and some words could possibly be identified if the listener was especially familiar with the dataset. As with GRID, the broad spectral detail appears to be estimated sufficiently well, as can be seen in the spectrograms of Figure 8.6, however the more fine spectral detail is lacking in the RM-3000 utterances. One benefit of the GRID grammar is that there are numerous different examples of words in the training data, whereas this is not the case for the larger RM-3000 corpus. The larger vocabulary size means an increased visual feature input-space, and more varied co-articulation effects as the grammar is considerably less constrained and more continuous, making the task more difficult. However, despite not yielding intelligible utterances, the spectral-envelope estimates produced for RM-3000 using the visual-to-audio mappings could still be usable within a multimodal speech enhancement or speaker separation system, as discussed in more detail in Section 9.2 on future work.

Despite lower intelligibility for RM-3000, it should be reinforced that the work presented in this thesis was developed for the GRID dataset, and, accordingly, further improvements could be made to increase the intelligibility of utterances from larger-vocabulary and less-constrained datasets given a more concentrated effort.

In the next chapter, conclusions are drawn for the work presented in this thesis on whether or not it is possible to reconstruct intelligible audio speech from visual speech information. Additionally, the limitations of this work are summarised with ideas for future work proposed on how these problems could be overcome.

# Chapter 9

# Conclusions

## 9.1 Summary and conclusions

The aim of this thesis has been to explore the visual-to-audio domain mapping problem for producing accurate spectral-envelope estimates, and to develop methods for generating suitable fundamental frequency and aperiodicity information, given input visual speech. The methods presented in this work were developed using speech from a male and female speaker from the GRID audiovisual corpus, within a speaker dependent configuration. Objective and subjective intelligibility evaluations of reconstructed utterances were performed to determine the performance of the various methods and configurations explored. To reconstruct audio speech, the STRAIGHT speech production model was chosen based on its successful application in speech synthesis and speech modification tasks. The model requires three parameters for speech reconstruction: a fundamental frequency contour, an aperiodicity surface, and a spectral-envelope surface.

To produce estimates of the spectral-envelope surface, visual-to-audio domain mapping models were developed. The mapping relies on the correlation that exists

between audio and visual speech, and, equivalently, between feature representations of the two modalities. Ultimately, Mel-filterbank audio features were selected for representing spectral-envelope, and AAM-based features were chosen for representing the visual articulators. Three broad methods were explored for performing the mapping: regression, clustering-and-classification with feature-level temporal encoding, and model-based methods of incorporating longer-range temporal information.

The first approach uses a regression system to produce estimates of real-valued, continuous audio feature vectors from single frames of visual input. Two mapping models are explored, GMMs and DNNs, with combinations of LPC and Mel-filterbank audio features and 2D-DCT and AAM visual features. The best system, using a GMM with LPC and AAM features, achieves audio-only subjective intelligibility scores of 40 %, which is a significant improvement over the intelligibility that would be achieved by chance alone, at 19 %. Furthermore, the final subjective intelligibility tests show a word-level accuracy of 30 % for the regression model using a DNN with Mel-filterbank and AAM features. These results provide initial evidence for the hypothesis that intelligible audio speech reconstructions can be generated from visual speech. However, that being said, the visual-only intelligibility was measured to be around 48 %, indicating that the audio reconstructions offered no benefit over the visual-modality alone. The next stage was to explore better methods of audio feature estimation.

The second approach uses a clustering-and-classification framework by reformulating the audio feature estimation problem as one of using classification models to estimate codebook entries of clustered audio features, with temporal information incorporated at the feature-level. Grouping windows of audio and visual feature vectors allowed for long-range temporal information to be incorporated, where a DNN is used to estimate windows of codebook entry labels from windows of AAM

visual features. Audio codebooks are produced from Mel-filterbank feature vectors using the mini-batch $k$-means algorithm, which allows for labels to be assigned to audio vectors for use in a classification model. The audio-only intelligibility for the best performing configuration, where window sizes of $S^A = S^V = 31$ are used, covering approximately 300 ms of audio and visual signal, is 77 %. The increase in performance over the regression approach is attributed to two main reasons. Firstly, that clustering of audio-features allows for more accurate spectral-envelope representations to be reconstructed, and secondly, that encoding longer-range temporal information, through clustering of grouped audio features on the order of 300 ms in length, allows for effects of co-articulation to be modelled.

In the third approach, temporal encoding is further investigated where longer-range dependencies in the data are modelled directly. Two methods are explored: using recurrent neural networks with the long-short term memory architecture, and using Viterbi decoding. The later method was found to perform poorly, with low objective intelligibility scores observed. The best performing model-level system, using the DB-LSTM with a sequence length of $T = 31$, achieves an audio-only intelligibility of 74 %. Confirming the results obtained for the feature-level approach, the results achieved for the DB-LSTM further indicates the importance of using audio and visual signal lengths of 300 ms for the visual-to-audio domain mapping models. This system showed considerably improved intelligibility over the regression approach, although the performance was slightly lower than the feature-level method.

For all three approaches, a further increase in intelligibility is observed by combining the reconstructed audio with the original video stream, with an average increase in word-level accuracy of 6–7 % for audiovisual media over audio-only. The highest intelligibility achieved was using utterances reconstructed from the feature-level approach combined with the original video, where a word-level accu-

racy of 84 % was observed. That is to say, for the female speaker chosen from the GRID audiovisual corpus, the average listener would correctly identify approximately 17 out of 20 words from utterances reconstructed using audio estimates from the best performing system. Accordingly, the main conclusion of this work is that intelligible audio speech can be reconstructed using information extracted solely from visual speech, for the given speaker from the GRID dataset.

To provide the excitation information as required by STRAIGHT, a number of approaches were explored for producing fundamental frequency contours and aperiodicity estimates. Three methods were presented for generating artificial-$f_0$ contours: unvoiced, monotone, and time-varying, where subjective tests established that the monotone method resulted in utterances with higher intelligibility over the other two methods. For aperiodicity estimation, two methods were proposed, where the approach using a joint clustering of spectral-envelope and aperiodicity information was found to give better estimates of the surface.

Lastly, experiments were conducted on a male speaker from the RM-3000 dataset using the best performing feature-level clustering-and-classification method recorded for the female speaker from the GRID corpus. Here, the aim was to determine whether the feature-level approach was able to yield intelligible audio speech reconstructions for a dataset with a larger-vocabulary and less-constrained speech. The lowest MSE recorded between the original and estimated Mel-filterbank features was obtained using an audio window of $S^A = 15$, covering 160 ms of audio signal, and a visual window of $S^V = 31$. Although the audio window is shorter than that used for GRID, it is again found that longer window widths, in comparison to using only a single frame, are important for visual-to-audio mapping. Objective intelligibility scores for reconstructed utterances of 0.612 and 1.693 were observed for STOI and PESQ respectively, whereas utterances reconstructed using the same method for the male speaker of the GRID corpus gave a STOI of 0.735

and PESQ of 2.055. These lower scores confirm informal listening tests which suggest that, whilst the signal exhibits broad speech-like characteristics, there is too much spectral-smoothing for words to be adequately identified. Furthermore, the spectral-envelope correlations between the original and estimated surfaces (see Figure 8.6) provide further evidence for the lower intelligibility of reconstructed utterances. Despite this, it is believed that given the promise shown by the work conducted on the GRID corpus, intelligible audio speech reconstructions could still be achieved for large-vocabulary datasets with less constrained grammars. Possible ideas to accomplish this are offered in the next section.

## 9.2 Future work

In this thesis, various approaches have been presented for using visual speech to produce the necessary parameters of a speech production model to give intelligible audio speech reconstructions. In this section, the potential limitations of the work are discussed, with suggestions for future work on generating better fundamental frequency contours, applying the techniques to speech enhancement and speaker separation systems, and improving intelligibility for larger datasets.

Methods were developed for producing artificial fundamental frequency contours as one of parameters required by STRAIGHT. The experiments conducted in the regression chapter (see Table 5.5) showed subjective intelligibility results for utterances reconstructed using spectral-envelope surfaces from two systems with original ground-truth $f_0$ contours and artificial excitation. For both systems, the intelligibility of utterances reconstructed using the original $f_0$, with correct voicing decisions, was around 10 % higher than when using unvoiced excitation for audio-only media. A similar increase in intelligibility of 9 % was observed for audiovisual media. Although fundamental frequency cannot be obtained directly from visual

speech, these results show the importance of using the correct $f_0$ contour, where speech intelligibility could be further improved by reconstructing utterances using more realistic contours and more accurate voicing decisions.

In experiments conducted by Shao and Milner [2004], an HMM is used to estimate $f_0$ values from MFCC vectors, where the optimum system shows a root mean squared (RMS) error of 3.1 Hz between the estimated and original fundamental frequency values. Accordingly, this technique could be explored for the work presented in this thesis by producing $f_0$ estimates from Mel-filterbank features estimated using a visual-to-audio mapping model. Furthermore, given the successful application of the DB-LSTM for sequential data in this work and others, this neural network architecture could also be compared with the HMM approach. As for the voicing decision, further experiments could be conducted on the voicing classification model presented in Chapter 4 to try and improve the accuracy. Voicing label estimates from a model with increased accuracy could then be combined with the $f_0$ contours and aperiodicity estimates to yield more faithful excitation information.

In Chapter 2, a number of systems using audio estimates from visual speech were reviewed for use in speech enhancement and speaker separation scenarios. The basic idea behind these systems was that degraded speech signals, either due to background noise or interfering speakers, could be cleaned by using the audio estimates within filtering and masking approaches. The majority of these systems use GMMs to jointly model the audio and visual features, with application of MMSE to produce audio estimates from input visual speech. This work has presented two broad approaches using deep neural network architectures that yield considerably improved audio estimates. Accordingly, it is believed that audiovisual speech enhancement and speaker separation systems applying these techniques to audio estimation would show increased performance.

Despite not yielding intelligible audio speech reconstructions on the larger RM-3000 dataset, where the utterances were less-constrained and taken from a larger vocabulary, the characteristics of the speech signal were similar to those exhibited for the female and male speakers from GRID for early developments of the regression system from Chapter 5. From informal listening tests it was found that the reconstructed signals were speech-like, and that upon listening to the original utterance and then its reconstructed counterpart, there was a noticeable similarity between the two. Furthermore, the recent improvement in state-of-the-art lip-reading systems, where word-level accuracies of 76 % are reported for RM-3000, suggest that intelligible audio speech could be reconstructed for larger datasets, and indicates a word-level accuracy to strive for.

To improve intelligibility for larger datasets, it would be beneficial to first design an audiovisual dataset with increasing vocabulary size—perhaps in increments of fifty words—and to collect a large amount of data for a single speaker. Experiments can then be conducted starting with the smallest vocabulary and increasing the size once the intelligibility has been maximised for that number of words. In this way, the methods developed showing good performance for a larger vocabulary would continue to work for smaller vocabularies. Furthermore, using increments in vocabulary size may perhaps suggest an upper-bound to the intelligibility that can be achieved.

## 9.3 Applications

In Chapter 1, two applications were proposed for utilising intelligible audio speech reconstructions from visual speech. The first was for surveillance scenarios where only video footage of a target is available, and the second was for development of a device for aiding laryngectomy patients with speech production. Applying this

work to these two applications is now discussed.

For surveillance scenarios, the visual-to-audio system could be incorporated into a software application for reconstructing audio speech from a video recording of a target speaker. It would not be required for the system to produce speech reconstructions in real-time as the emphasis would be more on obtaining accurate and reliable transcriptions, and so more effort could be spent on allowing for configurable parameters that would give the best chance of increasing intelligibility. Assuming that there is no training data available for the target, a number of pre-trained visual-to-audio models could be provided, where a user can select the model that gives the best performance. Furthermore, the fundamental frequency contours could be generated where the mean $f_0$ values could be chosen based on prior knowledge of the target's gender. Given training data, speaker adaptation techniques could be applied to produce speech reconstructions that better match that of the target speaker. As with silent speech interfaces, discussed next, this system would also benefit from visual-to-audio models that perform better on bigger datasets.

For laryngectomy patients, a device is envisaged that allows for real-time speech reconstructions for aiding conversational speech capabilities. Numerous medical devices are available for people with their larynx removed, including electrolarynx devices (artificial voice-box) and Permanent Magnet Articulography (PMA) systems, where each have their advantages and disadvantages. For example, although an electrolarynx allows for conversational speech, the quality is highly robotic and unnatural. To use PMA, an invasive procedure is required to place magnets in the tongue and lips of a patient to allow the device to be used normally.

A device using methods from this work would allow for hands-free speech communication with no operation required, at the expense of having a smaller vocab-

ulary, and is imagined as follows. A small form-factor video camera is located where a microphone is usually situated on a standard headset, and is focused on the mouth of the speaker. Visual features are extracted in real-time and passed to the visual-to-audio domain mapping models to produce spectral-envelope estimates which are subsequently used to reconstruct audio speech using a speech production model. This audio signal is then output using a speaker located about the person. Experiments could, therefore, be conducted on applying the visual-to-audio mapping techniques to function within a real-time scenario, and to determine an optimal vocabulary with which to allow for normal conversational speech.

# Appendix A

# Datasets

## A.1 GRID

The GRID audiovisual speech corpus collected by Cooke et al. [2006] contains low- and high-definition video and audio recordings of thirty-four speakers, of which 18 are male and 16 are female. The individual speakers can be seen in Figure A.1. For each speaker there are recordings of one thousand utterances each with a length of three seconds, giving 50 minutes of data in total. The ages of the speakers range from 18 to 49, with all but two of the speakers having British accents. Sentences take the form:

<command> <colour> <preposition> <letter> <digit> <adverb>,

and follow the grammar as displayed in Table A.1.

The video has a frame rate of twenty-five frames per second, giving seventy-five frames per three-second video. The high-resolution frame size is $720 \times 576$ pixels, and the low-resolution frame size is $360 \times 288$. Both sets of video contains full red-green-blue (RGB) colour information. Accompanying the dataset are word


**Figure A.1:** Stills from videos of each of the thirty-four speakers in the GRID audiovisual corpus. Speakers three and four are used for the experiments in this thesis.

time-alignment files for each utterance that describe the start and end points for each word, including periods of silence. Separately recorded audio, sampled at 50 kHz, accompanies the video stream. Furthermore, two sets of imaged-based 2D-DCT visual features are provided. One set contains features extracted from a region of interest that is stationary throughout the video, and the other from a region of interest located about a tracked point localised to the mouth of the speaker.

| command | colour | preposition | letter | digit | adverb |
|---|---|---|---|---|---|
| bin | blue | at | A-Z | 1-9 | again |
| lay | green | by | minus W | zero | now |
| place | red | in | | | please |
| set | white | with | | | soon |

**Table A.1:** GRID sentence grammar with available word choices for each of the six categories.

For the experiments conducted in this thesis, speakers three (male) and four (female) were used. From experiments presented in Cooke et al. [2006], the two speakers selected were found to have low word error rates in an automatic speech recognition task. Informal listening tests also suggested that their speech was highly intelligible in comparison to other speakers in the dataset.

## A.2 RM-3000

The RM-3000 audiovisual corpus was collected by Howell [2015] for performing confusion modelling for lip-reading, where it was found that other large-vocabulary audiovisual datasets contained too few data. The corpus contains 3000 utterances spoken by a native English male speaker, with sentences from the Resource Management (RM) corpus [Price et al., 1988]. The vocabulary contains 1000 words,

and lends itself well for continuous audiovisual speech processing applications. The sentence length ranges from 2–12 s, with an average of 5 s.

The video information was captured at 60 frames per second with a resolution of $1920 \times 1080$ pixels. The camera was placed in front of the speaker to record a full-frontal pose. A clip-on microphone was used to record the audio with a sampling frequency of 48 kHz. Pre-extracted AAM features (see Section 3.5) of the inner- and outer-lip are provided, having been extracted from the video re-sampled to a resolution of $640 \times 360$ pixels. The AAM visual feature vector dimensionality was chosen to retain 95 % of the shape variation, and 90 % of the appearance variation. Furthermore, phoneme transcriptions are provided. The training set is comprised of 2250 sentences, with the remaining 750 sentences used for testing.

# Appendix B

# Neural network architectures

## B.1 Introduction

To allow for interested readers of this thesis to repeat the experiments presented, this appendix details the various neural network architectures used, and includes information on any pertinent pre- and post-processing performed. As with any use of neural networks, a certain amount of trial and error is required to get the algorithms to function well. Accordingly, it would be possible to improve performance by trying alternative values and configurations from the ones described below.

The Python programming language was used for all software implementations. Data was processed using both the NumPy[1] and SciPy[2] libraries, which contain myriad functions for manipulating data. To construct and train the neural networks, the Lasagne[3] and theanets[4] libraries were used. Both of these libraries provide abstractions for the fantastic Theano[5] library — one of the dominant

[1] www.numpy.org
[2] www.scipy.org
[3] github.com/Lasagne/Lasagne
[4] github.com/lmjohns3/theanets
[5] deeplearning.net/software/theano/

toolkits used for producing CUDA code from mathematical expressions, allowing equations to be performed on GPUs.

## B.2 Excitation models

### B.2.1 Single-layer neural network

| | |
|---|---|
| Library | theanets |
| Data preprocessing | z-score normalisation is applied to visual features |
| Learning rate | 0.0001 |
| Regularisation | Dropout applied to hidden layers with probability p(0.5) |
| Batch size | 256 |
| Input layer | Dimensionality is dependent on how many frames of visual features are concatenated |
| Hidden layers | Three layers with 1024 units each, using ReLU activations |
| Output layer | Softmax function |
| Loss function | Categorical cross-entropy |
| Optimisation function | RMSProp |

### B.2.2 Convolutional neural network

| | |
|---|---|
| Library | Lasagne |

| | |
|---|---|
| Data preprocessing | Images converted to grayscale, and values rescaled to be between 0 and 1 |
| Learning rate | 0.001 initially, with annealing performed at 1% per epoch |
| Regularisation | Dropout is applied to the single dense layer with probability p(0.5) |
| Batch size | 64 |
| Input layer | Dimensionality is based on size of input images, $64 \times 64$ pixels for the mouth images |
| Hidden layers | Convolutional layer with 100 filters of size $5 \times 5$<br>Max-pooling layer with window of size $2 \times 2$<br>Convolutional layer with 100 filters of size $5 \times 5$<br>Max-pooling layer with window of size $2 \times 2$<br>Convolutional layer with 100 filters of size $3 \times 3$<br>Max-pooling layer with window of size $2 \times 2$<br>Fully-connected layer with 512 units |
| Output layer | Softmax function, 3 classes |
| Loss function | Categorical cross-entropy |
| Optimisation function | RMSProp |

## B.3 Regression system DNN

| | |
|---|---|
| Library | theanets |

| | |
|---|---|
| Data preprocessing | z-score normalisation is applied to visual and audio features |
| Learning rate | 0.001 |
| Regularisation | None |
| Batch size | 500 |
| Input layer | Dimensionality is dependent on how many frames of visual features are concatenated |
| Hidden layers | Three layers with 1024 units each, using ReLU activations |
| Output layer | Linear, number of outputs depends on the size of the audio feature codebook |
| Loss function | Categorical cross-entropy |
| Optimisation function | rprop |

## B.4 Classification system DNN

| | |
|---|---|
| Library | theanets |
| Data preprocessing | z-score normalisation is applied to visual features, followed by application of LDA using codebook entry location for class labels |
| Learning rate | 0.0001 |
| Regularisation | Dropout applied to hidden layers with probability p(0.5) |

| | |
|---|---|
| Batch size | 256 |
| Input layer | Dimensionality is dependent on how many frames of visual features are concatenated |
| Hidden layers | Three layers with 1024 units each, using ReLU activations |
| Output layer | Softmax function, number of outputs depends on the size of the audio feature codebook |
| Loss function | Categorical cross-entropy |
| Optimisation function | RMSProp |

## B.5 Model-level features DB-LSTM

| | |
|---|---|
| Library | theanets |
| Data preprocessing | z-score normalisation is applied to visual features, followed by application of LDA using codebook entry location for class labels |
| Learning rate | 0.0001 |
| Regularisation | Gaussian noise (with zero mean) is added to inputs with a weight of 0.6, any gradients above 1 are clipped |
| Batch size | 512 |
| Input layer | Same dimensionality as a single visual feature vector |

| | |
|---|---|
| Hidden layers | Three bi-directional layers with 500 units in total, where each layer is combined of a forward recurrent layer with 250 units, and a backwards recurrent layer with 250 units |
| Output layer | Softmax function, number of outputs depends on the size of the audio feature codebook |
| Loss function | Categorical cross-entropy |
| Optimisation function | RMSProp |

# Bibliography


Abdelaziz, A. H., Zeiler, S., and Kolossa, D. (2013). Twin-HMM-based audio-visual speech enhancement. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3726–3730. IEEE.

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 655–658. IEEE.

Allen, J. B. (1994). How do humans process and recognize speech? *Speech and Audio Processing, IEEE Transactions on*, 2(4):567–577.

Almajai, I., Cox, S., Harvey, R., and Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2722–2726. IEEE.

Almajai, I. and Milner, B. (2007). Maximising audio-visual speech correlation. In *AVSP*, page 17.

Almajai, I. and Milner, B. (2008). Using audio-visual features for robust voice activity detection in clean and noisy speech. In *Signal Processing Conference, 2008 16th European*, pages 1–5. IEEE.

Almajai, I. and Milner, B. (2011). Visually derived wiener filters for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1642–1651.

Almajai, I., Milner, B., and Darch, J. (2006). Analysis of correlation between audio and visual speech features for clean audio feature prediction in noise. In *INTERSPEECH*.

ANSI, A. (1997). S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*.

Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2016). Lipnet: Sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555.

Auer Jr, E. T. and Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, 102(6):3704–3710.

Baer, T., Moore, B. C., and Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *Journal of rehabilitation research and development*, 30:49–49.

Barker, J. P. and Berthommier, F. (1999). Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*.

Beerends, J. G., Larsen, E., Iyer, N., and van Vugt, J. M. (2004). Measurement of speech intelligibility based on the PESQ approach. *Measurement of Speech and Audio Quality in Networks (MESAQIN)*.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bernstein, L. (2012). Visual speech perception. *Audiovisual speech processing*, pages 21–39.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3):255–272.

Cairns, H. S. et al. (2010). *Fundamentals of psycholinguistics*. John Wiley & Sons.

Chen, B. Y., Zhu, Q., and Morgan, N. (2004). Learning long-term temporal features in LVCSR using neural networks. In *INTERSPEECH*.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2016). Lip reading sentences in the wild. *arXiv preprint arXiv:1611.05358*.

Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Cootes, T. F., Edwards, G. J., Taylor, C. J., et al. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

De Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Dean, D. and Sridharan, S. (2010). Dynamic visual features for audio–visual speaker verification. *Computer Speech & Language*, 24(2):136–149.

Deshmukh, O., Espy-Wilson, C. Y., Salomon, A., and Singh, J. (2005). Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):776–786.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope*, 58(9):955–991.

ETSI (2002). Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ES 202 050 version 1.1.1, ETSI STQ-Aurora DSR Working Group.

Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Interspeech*, pages 1964–1968.

Fletcher, H. (1953). Speech and hearing in communication.

Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4):377–430.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, 19(1):90–119.

Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P. K., and Garcia, O. N. (2005). Audio/visual mapping with cross-modal hidden markov models. *IEEE Transactions on Multimedia*, 7(2):243–252.

George, E. B. and Smith, M. J. (1997). Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406.

Girin, L., Schwartz, J.-L., and Feng, G. (2001). Audio-visual enhancement of speech in noise. *The Journal of the Acoustical Society of America*, 109(6):3007–3020.

Glotin, H., Vergyr, D., Neti, C., Potamianos, G., and Luettin, J. (2001). Weighting schemes for audio-visual fusion in speech recognition. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 173–176. IEEE.

Gonzalez, J. A., Green, P. D., Moore, R. K., Cheah, L. A., and Gilbert, J. M. (2015). A non-parametric articulatory-to-acoustic conversion system for silent speech using shared gaussian process dynamical models. In *Fifth Speech Conference of UK and Ireland*.

Gonzalez, S. and Brookes, M. (2014). PEFAC - A pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):518–530.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A. C., and Bengio, Y. (2013). Maxout networks. *ICML (3)*, 28:1319–1327.

Graves, A., Jaitly, N., and Mohamed, A.-r. (2013a). Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013b). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552.

Heiga, Z., Tomoki, T., Nakamura, M., and Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE transactions on information and systems*, 90(1):325–333.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Hiroya, S. and Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an hmm-based speech production model. *IEEE Transactions on Speech and Audio Processing*, 12(2):175–185.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. (2011). Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA. In *INTERSPEECH*, pages 3009–3012.

Holmes, J. and Holmes, W. (2001). *Speech synthesis and recognition*. CRC press.

Holube, I. and Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3):1703–1716.

Hong, P., Wen, Z., and Huang, T. S. (2002). Real-time speech-driven face animation with expressions using neural networks. *IEEE Transactions on neural networks*, 13(4):916–927.

House, A. S., Williams, C., Hecker, M. H., and Kryter, K. D. (1963). Psychoacoustic speech tests: A modified rhyme test. *The Journal of the Acoustical Society of America*, 35(11):1899–1899.

Howell, D. L. (2015). *Confusion Modelling for Lip-Reading*. PhD thesis, University of East Anglia.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE.

Jensen, J. and Hansen, J. H. (2001). Speech enhancement using a constrained iterative sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 9(7):731–740.

Kang, M. G. and Lee, B. G. (1988). A generalized vocal tract model for pole-zero type linear prediction. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP88*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, 117(4):2224–2237.

Kates, J. M. and Arehart, K. H. (2014). The hearing-aid speech perception index (HASPI). *Speech Communication*, 65:75–93.

Kato, A. (2017). *Hidden Markov model-based speech enhancement*. PhD thesis, University of East Anglia.

Katsamanis, A., Papandreou, G., and Maragos, P. (2009). Face active appearance modeling and speech acoustic information to recover articulation. *IEEE transactions on audio, speech, and language processing*, 17(3):411–422.

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1303–1306. IEEE.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207.

Kawahara, H. and Morise, M. (2011). Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework. *Sadhana*, 36(5):713–727.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3933–3936. IEEE.

Khan, F. and Milner, B. (2013). Speaker separation using visually-derived binary masks. In *AVSP*, pages 215–220.

Khan, F. and Milner, B. (2015). Using audio and visual information for single channel speaker separation. *Interspeech 2015*, pages 1517–1521.

Kitawaki, N., Nagabuchi, H., and Itoh, K. (1988). Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications*, 6(2):242–248.

Kleijn, W. B. and Paliwal, K. K. (1995). *Speech coding and synthesis*. Elsevier Science Inc.

Kressner, A. A., Anderson, D. V., and Rozell, C. J. (2013). Evaluating the generalization of the hearing aid speech quality index (HASQI). *IEEE transactions on audio, speech, and language processing*, 21(2):407–415.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lan, Y., Harvey, R., Theobald, B., Ong, E.-J., and Bowden, R. (2009). Comparing visual features for lipreading. In *International Conference on Auditory-Visual Speech Processing 2009*, pages 102–106.

Lan, Y., Harvey, R., and Theobald, B.-J. (2012). Insights into machine lip reading. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4825–4828. IEEE.

Laroche, J., Stylianou, Y., and Moulines, E. (1993). HNS: Speech modification based on a harmonic+noise model. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 550–553. IEEE.

Laures, J. S. and Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research*, 42(5):1148.

Le Cornu, T. and Milner, B. (2015). Voicing classification of visual speech using convolutional neural networks. In *FAAVSP-The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communications*, 28(1):84–95.

Ma, J., Hu, Y., and Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405.

Marques, J. S., Almeida, L. B., and Tribolet, J. M. (1990). Harmonic coding at 4.8 kb/s. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 17–20. IEEE.

Massaro, D. W., Beskow, J., Cohen, M. M., Fry, C. L., and Rodgriguez, T. (1999). Picture my voice: Audio to visual speech synthesis using artificial neural networks. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*.

McAulay, R. J. and Quatieri, T. (1995). Sinusoidal coding. In *Speech coding and synthesis*, pages 121–173. Elsevier Science Inc.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

Middelweerd, M. and Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, 82(6):2145–2147.

Miller, S. E., Schlauch, R. S., and Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noisea). *The Journal of the Acoustical Society of America*, 128(1):435–443.

Morales, S. O. C. and Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*, 2009(1):1–14.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological science*, 15(2):133–137.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

Nesterov, Y. et al. (2007). Gradient methods for minimizing composite objective function. Technical report, UCL.

Newman, J. L., Theobald, B.-J., and Cox, S. J. (2010). Limitations of visual speech recognition. In *AVSP*, page 1.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737.

Nye, P. and Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). *Haskins Laboratories Status Report on Speech Research, SR*, 33:77–91.

Ohtani, Y., Toda, T., Saruwatari, H., and Shikano, K. (2006). Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation.

O'Shaughnessy, D. (1988). Linear predictive coding. *Potentials, IEEE*, 7(1):29–32.

Paliwal, K. K. and Alsteris, L. D. (2003). Usefulness of phase spectrum in human speech perception. In *INTERSPEECH*.

Paliwal, K. K. and Atal, B. S. (1993). Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing*, 1(1):3–14.

Park, K.-Y. and Kim, H. S. (2000). Narrowband to wideband conversion of speech using GMM based transformation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1843–1846. IEEE.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Petajan, E. D. (1984). *Automatic lipreading to enhance speech recognition (speech reading)*. PhD thesis, University of Illinois at Urbana-Champaign.

Plomp, R. and Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18(1):43–52.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326.

Price, P., Fisher, W. M., Bernstein, J., and Pallett, D. S. (1988). The DARPA 1000-word resource management database for continuous speech recognition. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 651–654. IEEE.

Qian, Y., Fan, Y., Hu, W., and Soong, F. K. (2014). On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3829–3833. IEEE.

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (1988). *Objective measures of speech quality*. Prentice Hall.

Rabiner, L. and Juang, B.-H. (1993). Fundamentals of speech recognition.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice Hall.

Rao, K. R. and Yip, P. (2014). *Discrete cosine transform: algorithms, advantages, applications*. Academic press.

Richard, M. D. and Lippmann, R. P. (1991). Neural network classifiers estimate bayesian a posteriori probabilities. *Neural computation*, 3(4):461–483.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Neural Networks, 1993., IEEE International Conference On*, pages 586–591. IEEE.

Rivet, B., Wang, W., Naqvi, S. M., and Chambers, J. A. (2014). Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8614–8618. IEEE.

Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, pages 338–342.

Sayood, K. and Fow, E. (2000). Introduction to data compression. *India edition*.

Schmidt-Nielsen, A. (1992). Intelligibility and acceptability testing for speech technology. Technical report, DTIC Document.

Schroeder, M. and Atal, B. (1985). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 937–940. IEEE.

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2):B69–B78.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM.

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience research*, 47(3):277–287.

Shao, X. and Milner, B. (2004). Pitch prediction from mfcc vectors for speech reconstruction. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–97. IEEE.

Sharma, S., Ellis, D., Kajarekar, S., Jain, P., and Hermansky, H. (2000). Feature extraction using non-linear transformation for robust speech recognition on the aurora database. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II1117–II1120. IEEE.

Shi, G., Shanechi, M. M., and Aarabi, P. (2006). On the importance of phase in human speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1867–1874.

Silen, H., Helander, E., and Gabbouj, M. (2011). Prediction of voice aperiodicity based on spectral representations in hmm speech synthesis. In *Interspeech*, pages 105–108.

So, S. and Paliwal, K. K. (2007). A comparative study of LPC parameter representations and quantisation schemes for wideband speech coding. *Digital Signal Processing*, 17(1):114–137.

Spitzer, S. M., Liss, J. M., and Mattys, S. L. (2007). Acoustic cues to lexical segmentation: A study of resynthesized speech. *The Journal of the Acoustical Society of America*, 122(6):3678–3687.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Steeneken, H. J. (2001). The measurement of speech intelligibility. *Proceedings-Institute of Acoustics*, 23(8):69–76.

Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *Speech and Audio Processing, IEEE Transactions on*, 9(1):21–29.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R., editors, *Hearing by Eye: The Psychology of Lip-Reading*. Lawrence Erlbaum Associates.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 335(1273):71–78.

Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP*, pages 4214–4217.

Taylor, S., Kato, A., Matthews, I., and Milner, B. (2016). Audio-to-visual speech conversion using deep neural networks. *Interspeech 2016*, pages 1482–1486.

Ter Keurs, M., Festen, J. M., and Plomp, R. (1992). Effect of spectral envelope smearing on speech reception. i. *The Journal of the Acoustical Society of America*, 91(5):2872–2880.

Ter Keurs, M., Festen, J. M., and Plomp, R. (1993). Effect of spectral envelope smearing on speech reception. ii. *The Journal of the Acoustical Society of America*, 93(3):1547–1552.

Thangthai, K., Harvey, R., Cox, S., and Theobald, B.-J. (2015). Improving lipreading performance for robust audiovisual speech recognition using DNNs. In *Proc. FAAVSP, 1St Joint Conference on Facial Analysis, Animation and Audio–Visual Speech Processing*.

Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3):215–227.

Toda, T., Saruwatari, H., and Shikano, K. (2001). Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 841–844. IEEE.

Tu, Y., Du, J., Xu, Y., Dai, L., and Lee, C.-H. (2014). Deep neural network based speech separation for robust speech recognition. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 532–536. IEEE.

Valentini-Botinhao, C., Yamagishi, J., King, S., et al. (2011). Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise? In *Interspeech*, pages 1837–1840.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.

Voiers, W. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39.

Vos, K., Jensen, S., and Soerensen, K. (2010). Silk speech codec. *IETF draft*.

Wang, J., Shu, H., Zhang, L., Liu, Z., and Zhang, Y. (2013). The roles of fundamental frequency contours and sentence context in mandarin chinese speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):EL91–EL97.

Websdale, D., Le Cornu, T., and Milner, B. (2015). Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation. *Proceedings of Interspeech 2015*.

Weismer, G. (2008). Speech intelligibility. *The handbook of clinical linguistics*, pages 568–582.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., and Renals, S. (2009). Robust speaker-adaptive hmm-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1208–1230.

Yamagishi, J., Zen, H., Toda, T., and Tokuda, K. (2007). Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007. In *Proc. Blizzard Challenge 2007*.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1):23–43.

Zhou, Z., Zhao, G., Hong, X., and Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605.